SYLLABUS (5)

Data Mining and Business Intelligence - 2170715

1. Overview and concepts Data Warehousing and Business Intelligence

Why reporting and analysing data. Raw data to valuable information-Lifecycle of data - What is business intelligence - BI and DW in today's perspective - What is data warehousing - The building blocks : Defining features - Data warehouses and data marts - Overview of the components - Metadata in the data warehouse - Need for data warehousing - Basic elements of data warehousing - Trends in data warehousing. **(Chapter - 1)** 12%

2. The Architecture of BI and DW

BI and DW architectures and its types - Relation between BI and DW - OLAP (Online Analytical Processing) definitions - Difference between OLAP and OLTP - Dimensional analysis - What are cubes ? Drill-down and roll-up - Slice and dice or rotation - OLAP models - ROLAP versus MOLAP - Defining schemas : Stars, Snowflakes and fact constellations. **(Chapter - 2)** 15%

3. Introduction to Data Mining (DM) 8%

Motivation for data mining - Data mining-definition and functionalities - Classification of DM systems - DM task primitives - Integration of a data mining system with a database or a data warehouse - Issues in DM - KDD process. **(Chapter - 3)**

4. Data Pre-processing 16%

Why to pre-process data ? - Data cleaning : Missing values, Noisy data - Data integration and transformation - Data reduction : Data cube aggregation, Dimensionality reduction - Data compression - Numerosity reduction - Data mining primitives - Languages and system architectures : Task relevant data - Kind of knowledge to be mined - Discretization and concept hierarchy. **(Chapter - 4)**

5. Concept Description and Association Rule Mining 16%

What is concept description ? - Data generalization and summarization-based characterization - Attribute relevance - Class comparisons association rule mining : Market basket analysis - Basic concepts - Finding frequent item sets : Apriori algorithm - Generating rules - Improved Apriori algorithm - Incremental ARM - Associative classification - Rule mining. **(Chapter - 5)**

6. Classification and Prediction 16%

What is classification and prediction ? - Issues regarding classification and prediction :

- Classification methods : Decision tree, Bayesian classification, Rule based, CART, Neural network.
- Prediction methods : Linear and nonlinear regression, Logistic regression.

Introduction of tools such as DB Miner /WEKA/DTREG DM tools. **(Chapter - 6)**

7. Data Mining for Business Intelligence Applications 8%

Data mining for business applications like balanced scorecard, Fraud detection, Clickstream mining, Market segmentation, Retail industry, Telecommunications industry, Banking and finance and CRM etc.,

- Data analytics life cycle : Introduction to big data business analytics - State of the practice in analytics role of data scientists.
- Key roles for successful analytic project - Main phases of life cycle - Developing core deliverables for stakeholders. **(Chapter - 7)**

8. Advance Topics 8%

Introduction and basic concepts of following topics.

Clustering, Spatial mining, Web mining, Text mining.

Big data : Introduction to big data : Distributed file system - Big data and its importance, Four Vs, Drivers for big data, Big data analytics, Big data applications. Algorithms using map reduce, Matrix-Vector multiplication by map reduce. Introduction to Hadoop architecture : Hadoop architecture, Hadoop storage : HDFS, Common Hadoop shell commands, Anatomy of file write and read., NameNode, Secondary NameNode and DataNode, Hadoop mapreduce paradigm, Map and reduce tasks, Job, Task trackers - Cluster Setup - SSH and Hadoop configuration - HDFS administering - Monitoring and maintenance. **(Chapter - 8)**

History

Data :

Data are any facts, numbers or text can be processed by a computer, Today, organization are accumulating vast and growing amounts of data in different formats and different database.

- Operational or transactional data such as sales, cost, inventory, payroll and accounting
- No operational data such as industry sales, forcast data and Macro economic data.
- Meta data - Data about the data itself, such as logical database design or data dictonary definitions.

Information :

The patterns, associations or relationship among all this data can provide Information for example : Analysis of retail point of sale transaction data can yield information on which products are selling.

Knowledge :

Information can be converted into knowledge about historical patterns and future trends. For example : Summary information on retail supermarket sales can be analyzed. A manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Application area of data mining :

Since data mining is a young disciplin with wide and diverse application there is still a nontrivial gap between general principles of data mining and domain - specific, effective data mining tools for particular application that are related to data mining.

- Data mining for biomedical and DNA data Analysis.
- Data mining for financial data analysis.
- Data mining for the telecommunication industry.
- Data mining for webmining.
- Data mining for multimedia data.
- Data mining for traffic monitoring.
- Data mining for physical analysis.
- Data mining for change detection.
- Data mining for visual data.
- Data mining for audio data.
- Data mining for scientific data.
- Data mining for statistical data.

- Data mining for genetic linkage analysis.
- Data mining for engineering data.
- Data mining for text data.
- Data mining for social network.
- Data mining for pictorial data.
- Data mining for bioinformatics.
- Data mining for market basket analysis.
- Data mining for oil and gas industry.
- Data mining for relational database.

1.1 Why Reporting and Analyzing Data

Data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. However, algorithms and approaches may differ when applied to different type of data, Here are some examples in more detail :

i) Flat files : Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied the data in these files can be transaction, time-series data, scientific measurement.

ii) Relational database : A relational database consist of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key.

The most commonly used query language for relational database is SQL, which allows retrivel and manipulation of the data stored in the table as well as the calculation of aggregate function such as averages, sum, min, max and count.

For example :

Student :

Student-id	Name	Add	Age	Degree	Class	Branch	Other

Subject :

Student-code	S-name	S-aurthor	Publication	Type	Price	Univercity	Place	Other

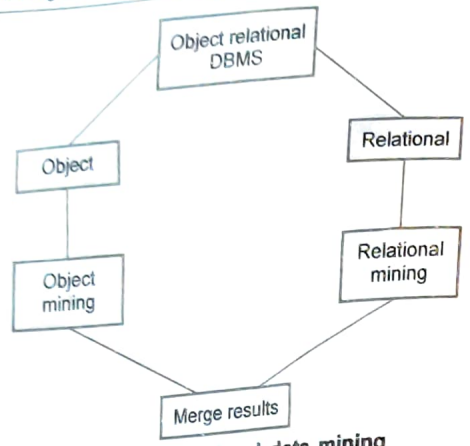


Fig. 1.1.1 Relational data mining

Data mining algorithms using relational database can be more varsatile than data mining algorithm specifically written for flat files. since they can take advantage of the structure inherent to relational database.

iii) Data warehouse : A data warehouse as a storehouse is a repository of data collected from multiple data sources (often heterogeneous) and is intended to be used as a whole under the some unified schema. A data warehouse gives the option to analyze data from different source under the same roof.

Let us suppose that our video store becomes a franchise in North America. Many video stores belonging to our video store company may have different database and different structures. If the executive of the company wants to access the data from all stores for strategic decision - making future direction, marketing, it would be more appropriate to store all the data in one site with a homogeneous structure that allows interactive analysis.

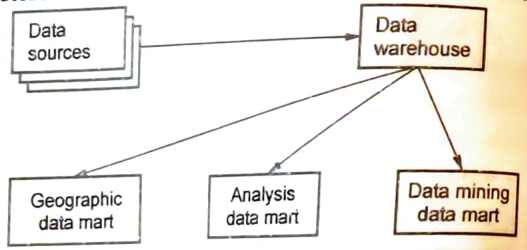


Fig. 1.1.2 Data warehouse and its relation with other streams

Data mining data mart extracted from a data warehouse a data warehouse is not a requirement for data mining. Setting a large data warehouse that consolidates data from multiple sources, resolves data integrity problems and loads the data into a query database can be an enormous task, sometimes taking years and costing millions of dollars, you could, however, mine data from one or more operational or transactional database try simply extracting it into a read - only database. This New database function as a type of data mart.

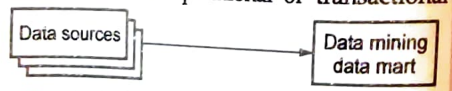


Fig. 1.1.3 Data warehouse and data mart

iv) Transaction database : A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of times. Associated with the transaction files could also be descriptive data for the items.

For example : In the case of the video store, the rentals table represents the transaction database. Each record is a rental contract with a customer identifier, a data and the list of items rented (videotapes, games, VCR etc). Since relation database do not allow nested tables (a set as attribute value) transaction are usually stored in flat files or stored in two normalized transaction tables, one for the transaction and one for the transaction items. One typical data mining analysis on such data is called Market basket analysis or association rules in which association between items occurring together or in sequence are studied.

Transaction - ID	List of item-IDs/Name
T 100	Video, game, sound, VCD
T 200	VCR, CD, play
T 300	game, VCD, CD
T 400	CD, VCR
T 500	VCD, CD

Fig. 1.1.4 Transaction database

v) Multimedia database : Multimedia database include video, images, audio and text media, they can be stored on extended object - relational or object - oriented database, or simply on a file system, multimedia is characterized by its high dimensionality, which makes data mining even more challenging. Data mining from multimedia repositories may require computer vision, computer graphics, image interpretation and natural language processing methodologies.

Such data are reffered to as continuous media data.

A. Heterogeneous database

B. Lagacy database

C. Data streams.

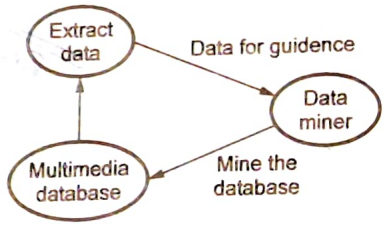


Fig. 1.1.5 Data for multimedia mining

vi) Spatial database : Spatial database are database that, in addition to usual data, store geographical information like maps and global or regional positioning. Such spatial database present new challenges to data mining algorithms.

For example : Include geographic (map) database, Very Large Scale Integration (VLSI) or computed - aided design database and medical and satellite image database. Spatial

data may be represented in raster format, consisting of n-dimensional bit map or pixel map.

vii) Time-series database : Time series database contain time related data such stock market data or logged activities. These database usually have a continuous flow of new data coming in which sometime cause the need for a challenging real time analysis. Data mining in such database commonly includes the study of trends and correlations between evolutions of different variable as well as the prediction of trends and movements of the variable in time.

viii) World wide web : The world-wide web and its associated distributed information services, such as yahoo, google, America online and Altavista provide rich worldwide, online information, services, were data objects are linked together to facilitate interactive access. Users seeking information of interest travers from one object via links to another, such systems provide sample opportunities and challenges for data mining.

1.2 Raw Data to Valuable Information-Life Cycle of Data

The kinds of patterns that can be discovered depend upon the data mining tasks.

- Descriptive data mining tasks describe the general properties of the existing data.
- Predictive data mining task that attempt to do predictions based on infrence on available data.
- Data warehousing explain ideal vision of properly maintaning a central repository of all organizational data.
- Data mining tool can be predict to understand the current market, future trends, behaviours and enables you to take proactive measure to more profit.
- Provide information to improve the business process.

1.2.1 Life Cycle of Data

Information about your organization from available data. Data life cycle having information data become information when you can use it to answer business question. So you can understand your business letter, data life cycle allow you to answer those question. So that dection makers at all level can respond quickly to change in the business. Providing answer to basic question such as,

- What are five top selling product ?
- How do sales this year ?
- What is moving average of sales ?
- Which place is best for our business ?
- Which product can be launch ?

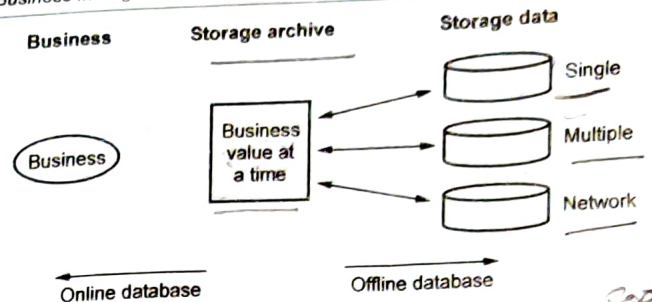


Fig. 1.2.1 Life cycle of data

1.3 What is Business Intelligence

Business intelligence or BI is the process of getting information about your business from available data.

In the information age, corporations have at their disposal massive amount of data, collected in transational system. These system are essential for business to keep track of their affairs.

Having data is not the same as having information, data becomes information when you can use it to answer business question, so you can understand your business letter, business intelligence allows you to answer those question, so that decision makers at all levels can respond quickly to changes in the business.

Business intelligence provides answers to basic question such as :

- What are five top selling products ?
- How do my sales this year ?
- What is mining average of my sales ?
- Which is best place of business ?
- Which product can be launch ?

1.4 BI and DW in Today's Perspective

Business intelligence has become a very imported activity in the business are an irrespective of the domain due to the fact that manager need to analyses comprehensively in order to face the challenges.

a) Data sourcing b) Data analysing, c) Data extracting.

Information correct for given criteria assessing the risks for supporting the decision making process in business intelligence perspective.

The key to **sourcing** is to be obtain the information in electronic form.

Example : Typically scanners, digital cameras, database queries, web searches, compute file access would be play significant role in a today's business intelligence perspective.

Data analysis granularity of the data to be extracted, possibility of data being extracted from identified sources and the confirmation that only correct and accurate data is extracted and passed on to data analysis.

Data mining is synthesizing useful knowledge from collections of data should be done in an analytical way using the in-depth. Business knowledge while estimating current trends, integrating and summarizing disparate information, validating models of understanding and predicting missing information or future trends. This process is called data mining process in BI perspective.

Example : Core stakeholders.

What is Data Warehousing

Data warehouse were developed in the late 1980's to meet growing demands for data analyzing and information management that could not be achieved by operational systems. Because the operational systems were designed in such as way that optimize for transactions only and number of operational or transaction system were growing quickly across departments inside an organization that make the data integration more difficult, this created problem of data redundancy, data integration, analysis and performance in reporting. As a result, a separated system call data warehouse is designed to solved those problems. Data warehouse system can bring data from various source systems such as relation data management system, flat files, spreadsheets, even remote data sources outside organization.

This data then is organized in such a way that optimized for enable business users and decision makers to access data in the form of useful information with ease of use.

Architecture of a Data Warehouse

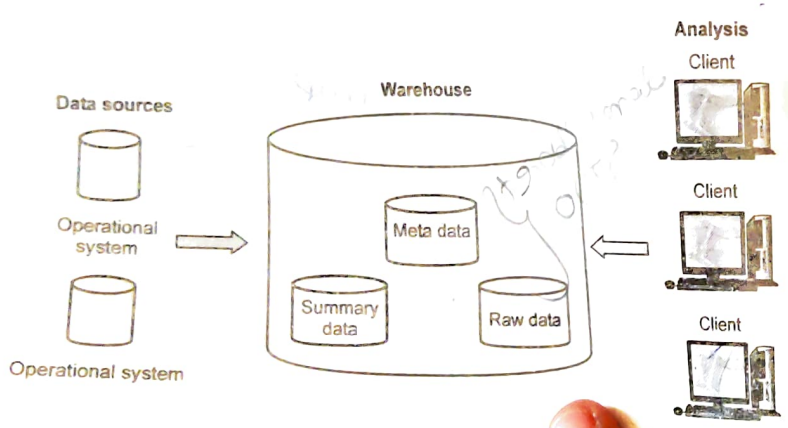


Fig. 1.5.1 Data warehous

Fig. 1.5.1 shown a simple architecture for a data warehouse. End user directly access data derived from several sources system through the data warehouse in Fig. 1.5.1 the meta data and row data of a traditional OLTP (Online Transaction Process) system is present as is an additional type of data, summary data, summaries are very valuable in data warehouses because they pre-computer long operation in advance.

1.5.1 Benefits of Data Warehouse

There are lot of benefits that data warehouse brings to organization.

- Keep history data for analyzing even if the source systems do not maintain historical data.
- Allow center point of accessing data across enterprises.
- Improve data quality by clearing and transfer mining data when loading it into data warehouse.
- Give business user or decision maker "a single varsion of truth" or information is presented consistently.
- Provide information instead of data to business user and decision makers.
- Provide oprimized query perfomance without impacting the operational systems.
- Provide information to improve the business processes.

1.5.2 Characteristics of Data Warehousing

"A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management" is decision making process.

i) Subject-oriented :

Data in an organization is organized in major objects or business process. The common example of subject oriented data are customer product, vendor and sale transaction.

ii) Integrated :

Integration is closely related to subject orientation. Data warehouse must put data from disparate source into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this goal, they are said to be integrated.

iii) Non-volatile :

A data warehouse is always a physically separate store of data transformed from the application data found in operational environment.

Non-volatile means that once entered into the warehouse, data should not change, this is logical because the purpose of a warehouse is to enable your to analyze what has occurred.

iv) Time-variant :

Data in data warehouse associates with time the time can be a single moment or span of time.

Support management's in decision making process the outcomes of a data warehouse are helping making decision based on historical data or facts. Then from business, to business process can be optimized to increase the efficiency and effectiveness.

1.6 The Building Blocks : Defining Features

a) Data warehouse :

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other source. It separates analysis workload from transaction workload and enables an organization to consolidate data from several source.

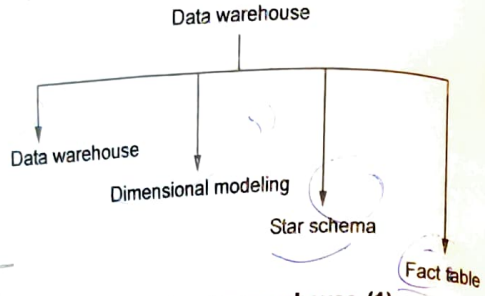


Fig. 1.6.1 Data warehouse (1)

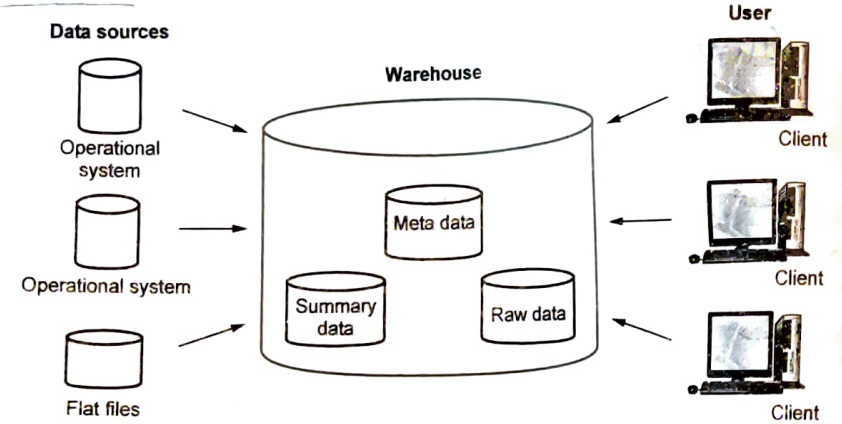


Fig. 1.6.2 Data warehouse (2)

- **Subject oriented :** Data in an organization is organized in major objects or business process. The common example of subject oriented data are customer product, vendor and sale transaction.
- **Integrated :** Integration is closely related to subject orientation. Data warehouse must put data from disparate source in to a consistent format. They must resolve

such problems as naming conflicts and inconsistencies among units of measure. When they achieve this goal, they are said to be integrated.

- **Non-volatile :** A data warehouse is always a physically separate store of data transformed from the application data found in operational environment.
- **Time variant :** Data in data warehouse associates with time the time can be a single moment or span of time.

b) Dimensional modeling :

Dimensional modeling is a database design technique to support business users to query data in data warehouse. The dimensional modeling is developed to be oriented around query performance and ease of use. It is important to note that the dimensional modeling is not necessary depends on relational database. The dimensional modeling handle approach is at logical level, which can be applied for any physical forms such as relational and multidimensional database.

In dimensional modeling, there are two important concepts.

i) Facts are also known as business measurement : Facts are normally numeric values which could be aggregated.

ii) Dimensions are called context : Dimension are business descriptors which specify the facts.

Process of Dimensional Modeling :

The following four-step process is commonly used in dimensional modeling design.

i) Select the business.
ii) Declare the grain.
iii) Identify the dimensions.
iv) Identify the fact.

c) Star schema :

Star schema is a dimensional design for a relational database often used in a data warehouse system. There is a fact table at the center of the schema surrounding by a number of dimension tables therefore the name star. Schema comes from appearance.

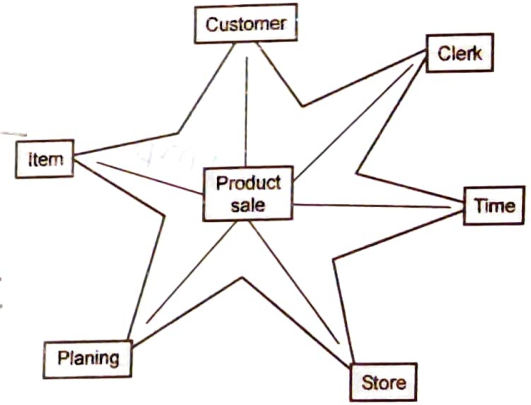


Fig. 1.6.3 Star schema

Star schema example :

- At the center of the schema we have a fact table called FACT-SALES. The primary key of the fact tables contains three surrogate keys associated with dimension table.
- Surrounding the fact table is number of dimension table Dim-Date, Dim-Store, Dim-Product.

Star schema :

- The more dimension table you add to the star schema, the more reporting possibilities it provides.
- The capability to study facts only depend on level of detail that the fact table stores.
- Star schema can help business analysis.

d) Fact table :

A fact table is usually in dimensional model in data warehouse design. It is often found at the center of a star schema surrounded by dimension table. Fact table consist of facts of a particular business process sale volume by month by product, fact are also known as measurement.

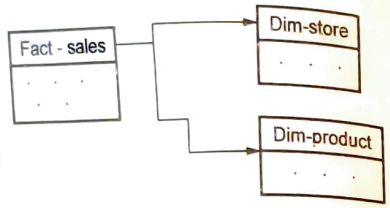


Fig. 1.6.4 Fact table

Different Type of Fact Table

i) Transnational

ii) Periodic

iii) Accumulating.

1.7 Data Warehouse and Data Marts

a) Data warehouse :

A data warehouse is one large data store for the business in concern which has integrated, time varient, non-volatile collection of data in support of managements decision making process. It will mainly have transactional data which would facilitate effective querying, analyzing and report generation which in turn would given the management the required level of information for the decision making.

Architecture of a data warehouse

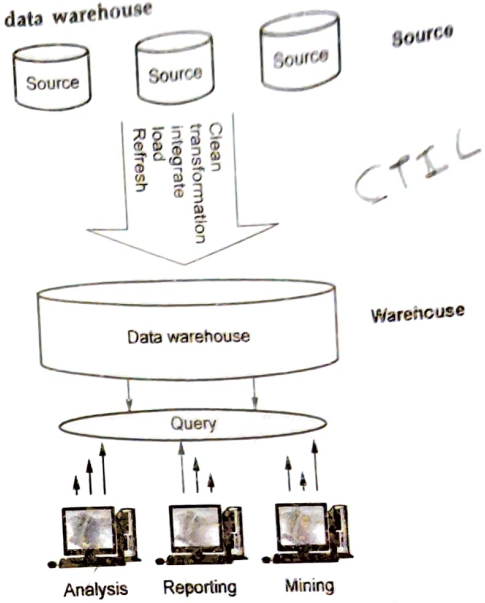


Fig. 1.7.1 Architecture of data warehouse

Why Selection of an Architecture of Data Warehouse ?

i) Information interdependence between organization units.

ii) Upper management information need.

iii) Nature of end user task.

iv) Constrains on resources.

v) Compatibility with existing system.

vi) Expert influence in system.

b) Data marts :

A data mart is a simple from of a data warehouse that is focused on a single subject such a sales, finance or marketing. Data marts are often built and controlled by a single department within an organization. Their single-subject focus, data marts usually draw data from only a few source. The source would be internal operational systems of data warehouse.

There are two type of data marts :

i) Independent or stand-alone data mart.

ii) Dependent data mart.

i) Independent or stand-alone data mart :

A stand-alone data mart focuses exclusively on one subject area and it is not designed in an enterprise context.

For example : Manufacturing has their data mart, human resources has their, finance has their and so on, stand alone data mart gets data from multiple transaction system in one subject area or department to support specific business needs. Stand alone data mart may use dimensional design or entity-relationship model.

ii) Dependent data mart :

Dependent data marts is some what simplified because formatted and summarized (clean) data has already been loaded into the central data warehouse.

There are several benefits of building a dependent data mart.

i) **Perforation :** When performance of a data warehouse becomes an issue, build one or two dependent data mart can solve problem.

ii) **Security :** By putting data outside data warehouse in dependent data mart, each department owns their data and has complete control over their data.

A data warehouse, unlike a data mart deals with multiple subject areas is typically implemented and controlled by a central organization unit such as the private sector (Information system) or Banking sector. It is called a central data warehouse.

Category	Data mart	Data warehouse
Finance	Business (Task)	Bank
Subject	Single	Multiple
Source	Minimum	Maximum
Manage Time	Minimum	Maximum
Use	Weakly/Monthly	Month to year
Security	easy	Difficult

1.7.1 Steps to Implemantation Data Mart

i) Designing ii) Constructing iii) Populating

iv) Accessing v) Managing.

i) Designing :

i) Identifying data source.

ii) For the business and technical requirement.

iii) For a particular organization purpose.

iv) Selecting a subset.

ii) Constructing :

i) Data processing

ii) Determine how best to set up the tables and the access structure.

iii) A physical data base and storage structure.

iii) Populating :

i) Extracting data.

ii) Ideal places for building and tracking.

iii) Data processing is performed outside.

iv) Accessing :

i) Set up an intermediate layer for the form-end tool to use.

ii) Manage business interface.

iii) Managing by using business tool.

v) Managing :

i) Optimizing the system for better performance.

ii) Manage centralize.

iii) A manage part of organization.

1.8 Virtual Warehouse

A virtual warehouse provides the opportunity for retailers to advertise a whole new line of items online that they would not otherwise have room for an their own shelyes. Through the distributors virtual warehouse services, customer can order product from the retailer website. The order is sent to the distributors warehouse, where it is picked, packed and shipped directly to the customer.

Benefits of virtual warehouse

i) Increasing customer loyalty through superior services.

ii) Virtual warehouse provide multiple ordering option and fast shipments.

iii) Appreciate a fully-stoked inventory.

iv) Virtual warehouse as opposed to a physical warehouse.

v) A virtual database is easy and fast.

vi) Is easy to build but require excess capacity on operational database server.

1.9 Meta Data in the Data Warehouse

Meta data is data about data that describes the data warehouse. It is used for building, maintaining, managing and using the data warehouse.

i) Technical meta data : Its contain information about warehouse data for use by warehouse designers and administrators.

ii) Business meta data : Its contain information easy-to-understand, perspective of the information can be stored in the data warehouse.

Meta data provide interactive access to users to help understand perspetive content and find data. One of the issues dealing with meta data relates to the fact that many data extraction tool capabilities to gather meta data remain immature. Meta data interface for user, which may involve some duplication effort.

Meta data architecture in data warehouse

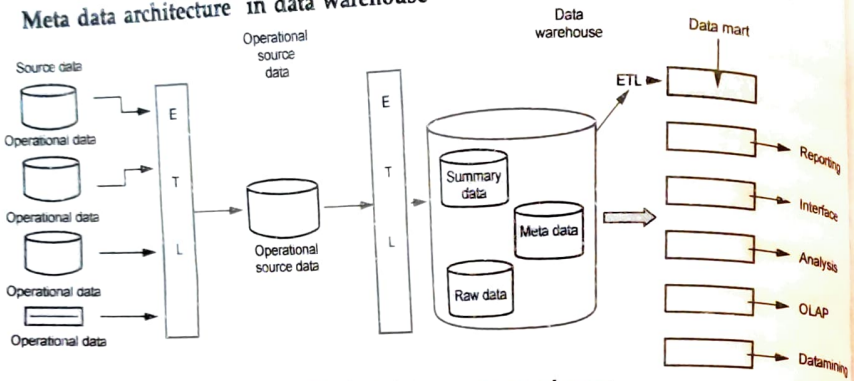


Fig. 1.9.1 Meta data in data warehouse

An important component of technical architecture is the staging process.

- Extract (E) : Data comes from multiple source and is of multiple type. Data compression and encryption handling must be considered at this area.
- Transform (T) : Data tranformation includes sarrogate key -management integration, cleaning and aggration.
- Load (L) : Loading is often multiple targets. (Optimize multiple target can be load)

1.9.1 Use of Meta Data in Warehouse

i) Source-to-target maps.

ii) Information on the contents of the data warehouse their location and structure.

iii) Job control (Job scheduling)

iv) Physical information.

v) Integrate and transformation.

1.9.1.1 Data Warehouse Manage to Meta Data

i) Model Management was built upon the result of having principled structure of data warehouse meta data. The early attempts in the area were largely based on the idea of mapping source and client schema to the data warehouse schema and tracing their attribute interdependencies.

ii) Data warehouse meta data with annotations concerning the quality of the collected data is quite large.

iii) Developers constructing or maintaining application, as well as the end-users interactively exploring the contents of the warehouse can benifit from the documentation facilities that data warehouse meta data.

iv) Managing use there are three broad categories of meta data

a) Build-time meta data b) Usage meta data c) Control meta data.

1.10 Need for Data Warehousing

i) Improving turnaround time for analysis and reporting.

ii) Data warehouse can work in conjunction with and hence; enhance the value of operational business application.

iii) Reducing cost to access historical data.

iv) Designing structures specifically to enable fast querying for business centric reporting.

v) Information in the data warehouse is under the control of data warehouse users so that even if the source system data is purged over time, the information in the warehouse can be stored safely for extended periods of time.

vi) Reimining informational processing load from transaction - oriented database.

vii) Integrating data from multiple process.

viii) Performing new type of analysis.

ix) Reduced administration.

x) Data warehouse to parallel performance.

1.10.1 Disadvantage of Data Warehouse

i) Data warehouse are not easy to maintain data.

ii) Data warehouse can have high costs.

1.10.1.1 Application of Data Warehouse

i) Insurance analysis. ii) Call report analysis.

iii) Share trading analysis. iv) E-Governance analysis.

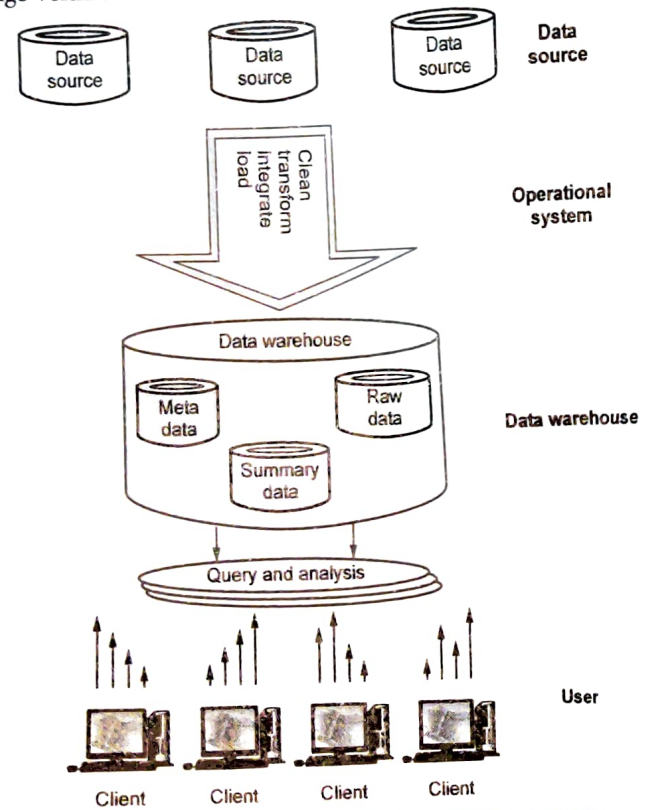
v) Banking analysis.

1.11 Basic Elements of Data Warehousing

A data warehouse is one large data store for the business in concern which is integrated, time variant, non-volatile collection of data in support of management's decision making process. It will mainly have transactional data which would facilitate effective querying, analyzing and report generation, which in turn would given the management the required level of information for the decision making.

Data Warehouse Architecture

i) The data must be loaded in the warehouse : The sheer volume of data in the warehouse makes loading the data a significant task. Monitoring tools for loads as well as methods to recover from incomplete or incorrect load are required with the huge volume of data in the warehouse.



1.11.1 Data warehouse architecture (Basic elements)

ii) Managing the warehouse : Developing and maintaining the infrastructure, data extraction, transformation and loading (ETL) process.

iii) Performance monitoring and measurement processes, quality management process and continuous improvement process.

iv) Data warehouse provide On-Line Analytical Processing (OLAP) for the interactive analysis of multidimensional data of varied granularities.

v) Provide information instead of data to business user and decision maker.

1.12 Recent Trends in Data Warehousing

Recent trends in many areas it evolves.

i) Implemented application show usage to expansion from tactical and data mart solution to strategic and enterprise data warehouse.

ii) Data warehouse initiatives continue to face fresh challenges that evolve with the changing business and technology environment.

iii) Data warehouse have developed new and more sophisticated technologies and have acquired and merged over.

iv) The awareness that unstructured data belongs in the data warehouse.

v) Data can back up incrementaly in warehouse.

vi) Data warehouse are designed to accommodate Ad-hoc queries.

vii) Data warehouse usually store or support historical analysis.

viii) Data warehouse provide retrieval of data without slow process system.

ix) Data warehousing with specific loading requirement.

x) Data warehouse requring data encryption technique.

1.12.1 Major Step of Recend Trends in Data Warehouse

Data warehouse follow up following recend trends.

i) Data warehouse administraction system.

ii) Data storage approach system.

iii) Configration improve in data warehouse system.

iv) DBMS in warehouse system.

v) Operational data warehouse system.

i) Data warehouse admistration system :

Data warehouse is being called on to support new initiatives, such as customer relationship management and supply chain management.

ii) Data storage apporach system :

User requirement and data realities drive the design of the dimensional approach. For example : customer name, product number, order slips, order date and sales person.

iii) Configration improve in data warehouse system :

Data warehouse is develop to making similar meaning data consistent when they are stored in the data warehouse the information deliver component is used to enable the process of subscribing for data warehouse information and having it delivered to one or more destination to user requirement.

iv) DBMS in warehouse system :

Designed for analysis of business measure by categories and attributes always implemented on the relational database management system. These approaches is as include in warehouse system.

i) Optimized for bulk load and large database size, Ad-hoc query processing and the need for flexible user view.

ii) Improve parallel relational database.

v) Operational data warehouse system :

Data warehouse updata every time an operational system performs, for a decision making purpose.

1.12.2 Characteristics of Data Warehouse

The characteristics of data warehouse are as follows :

i) Expert influence.

ii) Data extraction.

iii) Data transformation.

iv) Top-down and Bottem-up design.

v) Multi use support.

vi) Hybrid methodologies.

vii) Aggregation.

viii) Mapping.

ix) BI and transaction data.

x) Parallel DML.

1.12.3 Dynamic Warehouse (Dynamic Data Warehouse)

Data warehouse contain very large amount of data for multiple source in database or warehouse but its very difficult to manage it warehouse, at update autometaically every time an operational system performs a transaction is called dynamic warehouse this stage data warehouse generate transaction that are passed back in to operation system.

Dynamic warehouse up to data for critical manage data to supporting cost reduction initiative and open source business intelligence database.

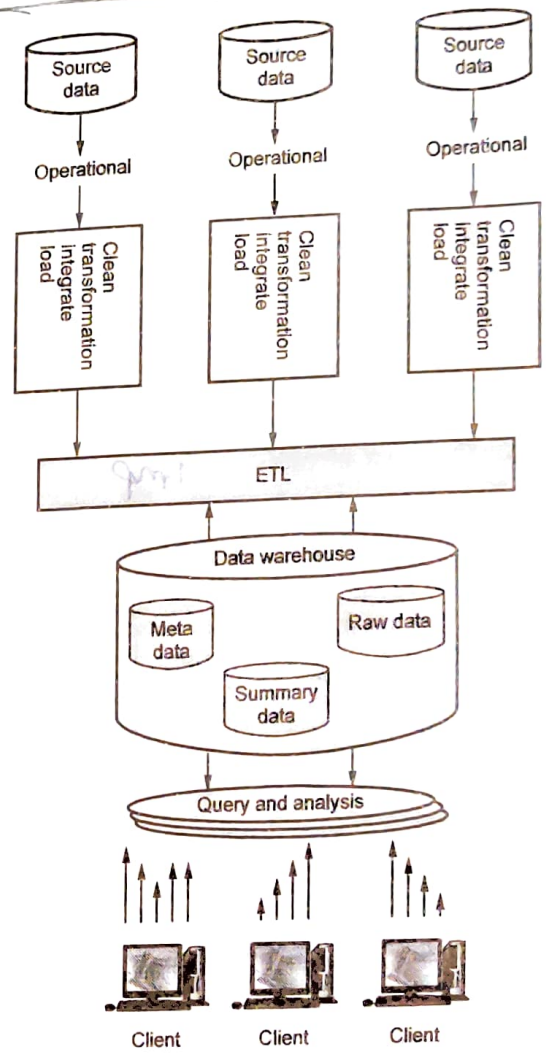


Fig. 1.12.1 Dynamic warehouse

In challenging time good decision making beocmes critical the best decision are made when all the relevent data available is taken into consideration. As users interactions with the dynamic warehouse increase, their approach to reviewing the results of their request for information can be expected to evolve from relatively simple manual analysis based on user defined thersholds configuration parameters and the information directory indicating where the appropriate source for the information can be found are all stored in warehouse.

Dynamic warehouseing application that support business process validated, reformatted, reorganized and summarized restructured with data from other source. The resulting data in dynamic warehouse becomes the main source of information for analysis and presentation through tool and query analysis. Benefit of dynamic warehouse.

i) **Real time updating :** Now often is the dynamic warehouse.

ii) **Drill down capabilities :** Dynamic warehouse makes it practical to use this capacity as much as needed.

iii) **Time to value :** How long does it take to get in dynamic warehouse.

iv) **Accuracy :** Dynamic warehouse components. Jointly categorized as the data query.

v) **Increased parallelism :** Dynamic warehouse can process queries involving in parallel.

1.12.4 3-tier Data Warehousing Architecture

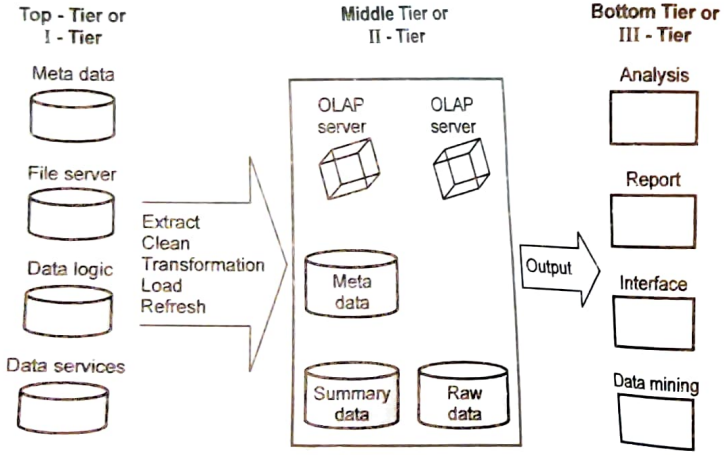


Fig. 1.12.2 3 - Tier data warehouse architecture

i) Bottom Tier → Database (RDBMS)

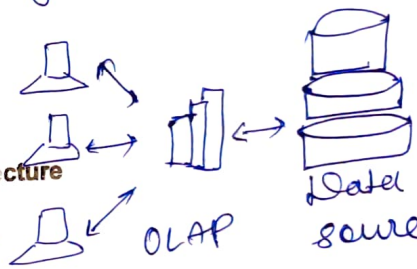
1. Bottom tier is a warehouse database server.
2. It's always DBMS relation. → after cleaning, transformation data is loaded.
3. Handling querying and materialization.
4. That is data partitioning.
5. It is a back-end tools.

ii) Middle-Tier → OLAP server

1. It is a relational in OLAP server (ROLAP).
2. Extended relational DBMS operation on multidimensional. → Business logic is there.
3. Hybrid OLAP (HOLAP) : Its user flexibility.
4. Multidimensional OLAP (MOLAP) : Special purpose.
5. OLAP servers is support snowflake.

iii) Top-Tier → Fevent end client layer.

1. It is a front-end client layer. → (Query tools)
2. Query reporting.
3. Data mining tool.
4. Client query.
5. Interface between middle tier.



1.12.5 Advantages of 3-tier Architecture

1. It can be makes logical separation.
2. Performance is very high.
3. Develop business logical layer.
4. It can be solve more complex problem.

Review Questions

1. *Define data mining.*
2. *What is data warehouse ?*
3. *Why need for data warehouse and its beniefits ?*
4. *What is the characteristic of data warehousing.*
5. *Explain following*
   i) Virtual warehouse ii) Use of meta data in warehouse iii) Dynamic warehouse

2 The Architecture of BI and DW

Syllabus

BI and DW architectures and its types - Relation between BI and DW - OLAP (Online Analytical Processing) definitions - Difference between OLAP and OLTP - Dimensional analysis - What are cubes ? Drill-down and roll-up - Slice and dice or rotation - OLAP models - ROLAP versus MOLAP - Defining schemas : Stars, Snowflakes and fact constellations.

Contents

2.1 What is Business Intelligence ?

Business Intelligence (BI) is the process, technologies and use tools to getting information from basic available data to develop knowledge into plans that drive profitable business action.

Business intelligence is not a new idea to develop business. It provides the ability to decision makers at all levels, can respond quickly to changes in the business, allows you to analyze current and long term trends, alerts you instantly to opportunities and problem and gives you continuous feedback on the effectiveness of your better decision. The following are key characteristics of BI,

i) Analytical skills

ii) Positive attitude

iii) Increased services

iv) Long term profitability

v) Responsiveness

2.1.1 Business Intelligence and Data Warehouse Architecture

BI architecture taking into consideration the value and quality of data as well as information flow in the system. The five layers and data source, ETL (Extract - Transform - Load), data warehouse, end user, and metadata layers. The rest of this section describes each of the layers. (See Fig. 2.1.1 on next page.)

Data source Layer :

In this layer there are basically two types, **internal and external.** In internal data source refers to data that is captured and maintained by operational systems inside an organization such as Customer Relationship Management and Enterprise Resource Planning Systems. Internal data sources include the data related to business operations (example, customers, products, and sales data).

These operational systems are also known as online transaction processing systems because they process large amount of transactions in real time and update data wherever it is needed. Operational systems contain only current data that is used to support daily business operations of an organization.

External data source refers to those that originate outside an organization. This type of data can be collected from external sources such as business partners, syndicate data suppliers, the Internet, Government and market research organizations.

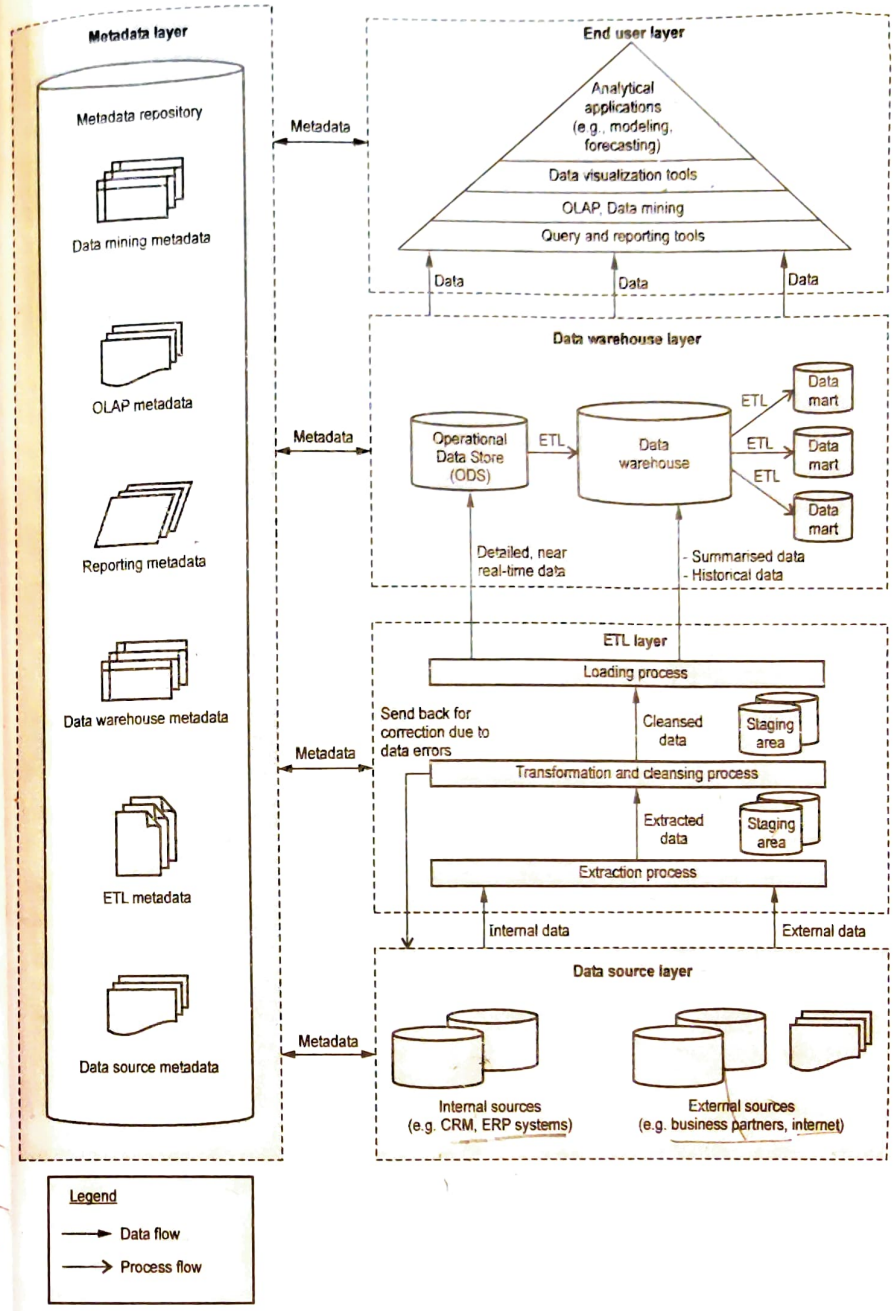


Fig. 2.1.1 Business intelligence data warehouse architecture

ETL (Extract - Transform - Load) Layer

Extraction is the process of identifying and collecting relevant data from different sources usually, the data collected from internal and external sources are not integrated, incomplete, and may be duplicated. Therefore, the extraction process is needed to select data that are significant in supporting organizational decision making.

The extracted data are then sent to a temporary storage area called the data staging area prior to the transformation and cleansing process. This is done to avoid the need of extracting data again should any problem occur. After that, the data will go through the transformation and the cleansing process. Transformation is the process of converting data using a set of business rules into consistent formats for reporting and analysis. Data transformation process also includes defining business logic for data mapping and standardizing data definitions in order to ensure consistency across an organization. As for data cleansing, it refers to the process of identifying and correcting data errors based on pre - specified rules. If there is an error found on the extracted data, then it is sent back to the data source for correction. Once data have been transformed and cleansed, they are stored in the staging area. This can prevent the need of transforming data again if the loading processes fail or terminate. Loading is the last phase of the ETL process. The data in staging area are loaded into target respository.

Data Warehouse Layer

There are three components in the data warehouse layer, namely operational data store, data warehouse, and data marts. Data flows from operational data store to data warehouse and subsequently to data mart.

Operational Data Store

The data stored in Operational Data Store (ODS) is volatile, which means it can be over - written or updated with new Data that Blow into ODS As such, ODS does not store any historical data. Generally, ODS is designed to support operational processing and reporting needs of a specific application by providing an integrated view of data across many different business applications. It is normally used by middle management level for daily management and short - term decision making.

Data Warehouse

The primary concept of data warehousing is that the data stored for business analysis can most effectively be accessed by separating it from the data in the operational systems. Many of the reasons for this separation have evolved over the years. In the past, legacy systems archived data onto tapes as it become inactive and many analysis reports ran from these tapes or mirror data sources to minimize the performance impact on the operational systems. These reasons to separate the operational data from analysis data have not significantly changed with the evolution of the data warehousing systems,

except that now they are considered more formally during the data warehouse building process.

Subject - oriented : Data from various sources are organized into groups based on common subject areas that an organization would like to focus on, such as customers, sales, and products.

Integrated : Data warehouse gathers data from various sources. All of these data must be consistent in terms of naming conventions, formats, and other related characteristics.

Time - variant : Each data stored in the data warehouse has time dimension to keep track of the changes or trends on the data. In other words, data warehouse will store historical changes on each piece of data.

Non - volatile : New data can be added into data warehouse regularly. But all the data stored in data warehouse are read - only. This means users are not allowed to update, over - write or delete the stored data.

Data Mart

While the data in a data warehouse is mainly used to support various needs across the whole organization, it is not equipped to support the needs and requirements of specific departments. Consequently, it is necessary to have data marts to support them, A data mart is subset of the data warehouse that is used to support analytical needs of a particular business function.

Metadata Layer

Metadata refers to data about data. It describes where data are being used and stored, the source of data, what changes have been made to the data, and how one piece of data relates to other information. Metadata repository is used to store technical and business information about data as well as business rules and data definitions. Good management and use of metadata can reduce development time, simplify on-going maintenance, and provide users with information about data. There are many different types of metadata to support at BI architecture such as data source, ETL, reporting, OLAP, and data mining metadata. Data source metadata consists of information about access mode, structure of data.

End User Layer

The end user layer consists of tools that display information in different formats to different users. These tools can be grouped hierarchically in a pyramid shape. As one moves from the bottom to the top of the pyramid, the degree of comprehensiveness at which data are being processed and presented increases.

Query and Reporting Tools

Query and reporting tools are very useful tools which allow end users to access and query data quickly, and to produce reports for decision making and management purposes. There are many different types of reports including standard reports, ad-hoc reports, budgeting and planning reports, and metadata reports. Both internal and external users can manage reports and other information easier and faster through BI portals. BI portal is a popular end user tool to deliver information.

OLAP (Online Analytical Processing)

One more OLAP servers can manage data in the data warehouse layer for reporting analysis, modeling, and planning to optimize business. OLAP server is a "data manipulation engine that is designed to support multidimensional data structures"

Roll-up or dril-up : It increases the level of aggregation, either by moving up to a higher level along dimensional hierachy or by reducing one or more dimensions from a given data cube.

Drill-down : It is the opposite of roll-up. It decreases the level of aggregation by moving down to a lower level (less detailed data) along a dimensional hierarchy or by adding one or more dimensions to a data cube.

Slice and dice : The slice operation can be performed by selecting a specific value on a single dimension, resulting in a sub-cube. The dice operation performs a projections on a data cube by selecting a range of values on two or more dimensions.

Pivot : It enables users to rotate the axes of the data cube, meaning swapping the dimensions to get different views of data.

Data Mining

Data mining process can be achieved with the integration of data warehouses and OLAP servers by performing further data analysis in OLAP cubes.

Data Visualization Tools

Data visualization tools such as dashboard and scorecards can be provided to managers and executives who need an overall view of their business performance. Dashboard is a useful tool that allows users to visualize data using charts, colored metrics or tables.

Analytical Applications

Applications that are equipped with analytical capabilities allow users to gain insights into improving the performance of business operations. By employing analytical applications, decision makers can also identify and understand what factors drive their business value, and thus able to leverage opportunities faster than their competitors.

2.2 Business Intelligence Development

Business intelligence development studio is a set of tools designed for creating business intelligences projects. Because business intelligence development studio was created as an IDE (Intelligence Development environment) in which you can create a complete solution, you work disconnected from the server.

You can change your data mining object as much as you want but the changes are not reflected on the server until after you deploy the project.

Working in an IDE is beneficial for the following reason.

- You have powerful customization tools available to configure business intelligence development studio to suit your needs.
- You can integrate your analysis services project with a variety of other business intelligence project encapsulating your entire solution into a single view.
- Full source control integration enables your entire into collaborate creating a complete business intelligence solution.

2.2.1 Working of Data Mining

Data mining gives you access to the information that you need to make intelligent decisions about business problem. Create a model that describes your business problem and then you run your data through an algorithm that generates a mathematical model of the data a process that is known as training the model.

2.2.2 Data Mining Concepts

Data mining is frequently described as "the process of extracting valid, authentic and actionable information form large database". Data mining derives pattern and trends that exist in data. These pattern and trends can be collected together and defined as a mining model. Mining models can be applied to specific business scenarios.

- Forecasting sales
- Targeting mailings towards specific customers.
- Determining which products are likely to be sold together.
- Finding sequence in the order that customer add product to a shopping cart.

2.3 Relation between BI and Data Mining

Business intelligence has become increasingly popular over the years and is currently a hot topic among many companies around the world. BI (Business intelligence) is often considered by companies to be tool for tuning their way of doing business by guiding their decision making business-wise in this way the individual company can make more

profitable decision based on intelligent analysis of their data depots. The main reason for using BI among companies is probably to increase profitability. Why use data depots for storage only when important and profitable market knowledge can be extracted from them using BI.

From a technical perspective making profit is not the only reason for using BI. Maintaining system and structure in large multidimentional data depots has always been an important task among with being able to analyze the contents of the depots. This task will in future become even more challenging because of the evolution of storage devices and processor power.

Data mining while the query-reporting_analysis is able to provide answer for question of the "what happened" kind. Data Mining Utilizes clever algorithms for a much deeper and intelligent analysis of data. BI solution using data mining techniques are then capable of handling "What will happen" and "How/why did this happen" matters. All this is done in a semi or full automatic process saving both time and resource.

2.4 Architecting Principles for Business Intelligence

The principal objective of business intelligence can be summed up as follows :

- To provide a single version of the truth across an entire organization.
- To provide a simplified system implementation, deployment and administration.
- To deliver strategic, tactical and operational knowledge and actionable insight.

Because of the focus on information in business intelligence application, the privileged point of view of the supporting architecture has to the information view from this point of view, the most paradigms are -

- The hub-and-spoke architecture with centralized data warehouse and dependant data marts.
- The data-mart bus architecture with linked con-formed dimensional data marts.
- Independent non-integrated data marts.

2.5 OLAP (Online Analytical Processing)

OLAP is a technology that is used to organize large business data base and support business intelligence. Data analytical processing for decision making technique with navigation such as summarization, dimension, consistant interactive access to a wide-veriety of possible views of information that has been transformation from raw data to reflect the real dimensionally of the enterprise as understood by user.

Online analytical processing adopt either a star diagram, flow chart dimensional model, for object oriented and time varient database purpose. These provide preprocessing functionalities such as the following

i) **Roll-up :** Data is summarized with increasing generalization.

ii) **Drill-down :** Increasing levels of details are revealed.

iii) **Pivot :** Cross tabulation that is rotation.

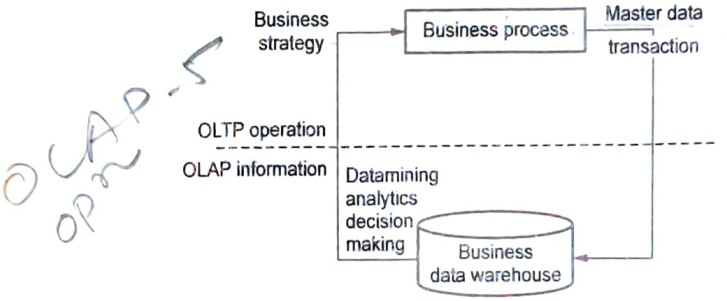


Fig. 2.5.1 OLAP

iv) **Drill-with :** Classification different switching one within the some dimension.

v) **Sorting :** Data is sorted by ordinal value.

vi) **Slice and dice :** Performing projection operation on the dimension.

vii) **Attributes :** Attribute are computed by operation on stored and derived values.

2.5.1 OLAP Server

The OLAP server is physically stage the processed multi-dimensional information to deliver consistent and rapid response times to end user it may populate its data structures is real-time from relational database.

OLAP server is a high-capacity, multi-user data manipulation engine specifically designed to support and operate on multi-dimensional data structures a multi-user server mode and offers consistently rapid response to queries, regardless of data base size and complexity. Implementation of a warehouse server for OLAP processsing include the following.

i) **OLTP server (On-line transaction process) :** A data warehouse support an OLTP system by providing a place for the OLTP database to offload data as it accumulates by providing services that would complicate and degrade OLTP operation if they were performed in the OLTP database.

ii) **MOLAP server (Multidimension on-line analytical process)** : A data warehouse provide a multi-dimensional view of data in an intutive model designed to match the type of queries posed by analysis and decision markers.

iii) **ROLAP server (Relational on-line analytical process)** : ROLAP servers include optimization for DBMS servers, to store and manage warehouse data.

2.5.2 Difference between OLTP and OLAP

Operation	OLTP system (Online Transaction Procession)	OLAP system (Online Analytical Processing)
Source of data	Operational data : OLTP are the original source of the data.	Consolidation data : OLAP data comes from the various OLTP database.
Purpose of data	To control and run-fundamental business tasks.	To help with planning problem solving and decision support.
Queries	Relatiely standaredized and simple queries returning relatively few records.	Often complex queries involving aggregation.
Database design	Highly normalized with many tables.	Typically due to the existence of aggregation structures and history data require more index than OLTP.
Data	Revels a snapshot of on going business process of data.	Multidimensional views of various kinds of business activities to involve.
Work	Simple transaction	Complex query
User	1000	100

2.5.2.1 Advantage of OLAP

- OLAP is performs intricate resolution of business data and provides the efficacy for complex calculation.
- Now OLAP creates a single platform for all the information and business needs to planning, budgeting for reporting and analysis.
- Intricate presentation can create an understanding of relationship not previously realized.

2.5.3 What are Cubes ?

OLAP cube is a structure that allows fast analysis of data according to the intricate dimensions that is eassy define a business problem.

OLAP Cube Advantages :

- OLAP cube is a technology that is stores data in optimized way to provide a quick response to various type of complex by using dimensions and measures.
- OLAP cube pre-aggregate the measures by the different levels of categories in the dimension to enable the quick response time.
- There are main reason for adding a cube to business solution.

1) **Drill down :** When this tool will automatically allow drilling up and down an dimension with the data source is an OLAP cube

2) **Performance :** OLAP cube structure and pre - aggregation allows to provide assay and fast response to queries that would have required grouping and summarizing

2.5.3.1 Roll - up

Roll - up performs aggregation on a data cube in any of the following ways :

- By climbing up a concept hierarchy for a dimension
- Fig. 2.5.2 illustrates how roll-up works. (Refer Fig. 2.5.2 on next page.)
- Roll-up is performed by climbing up a concept hierarchy for the dimension in location.
- Initially the concept hierarchy was "street < city < province < country".
- On rolling up, the data is aggregated by ascending the location hierarchy from the level of city to the level of country.
- The data is ground into cities rather than the countries.
- When roll-up is performed, one or more dimensions from the data cube are removed.

2.5.3.2 Drill-down

Drill-down is the reverse operation of roll-up. It is performed by either of the following ways :

- By stepping down a concept hierarchy for a dimension
- Fig. 2.5.3 illustrates how drill-down works (Refer Fig. 2.5.3 on page 2-13.)
- Drill-down is performed by stepping down a concept hierarchy for the dimension in time.
- Initially the concept hierarchy was "day < month < quarter < year."

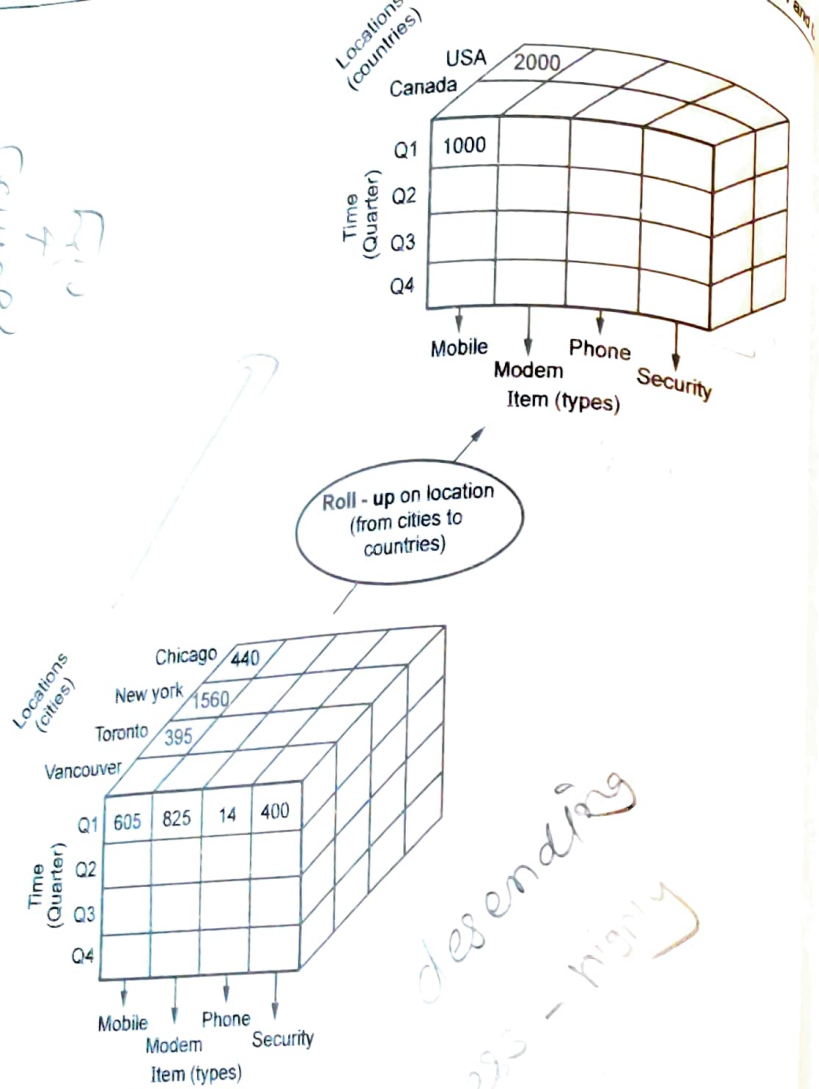


Fig. 2.5.2 Roll-up

- On drilling down, the time dimension is descended from the level of quarter to the level of month.
- When drill-down is performed, one or more dimensions from the data cube are added.
- It navigates the data from less detailed data to highly detailed data.

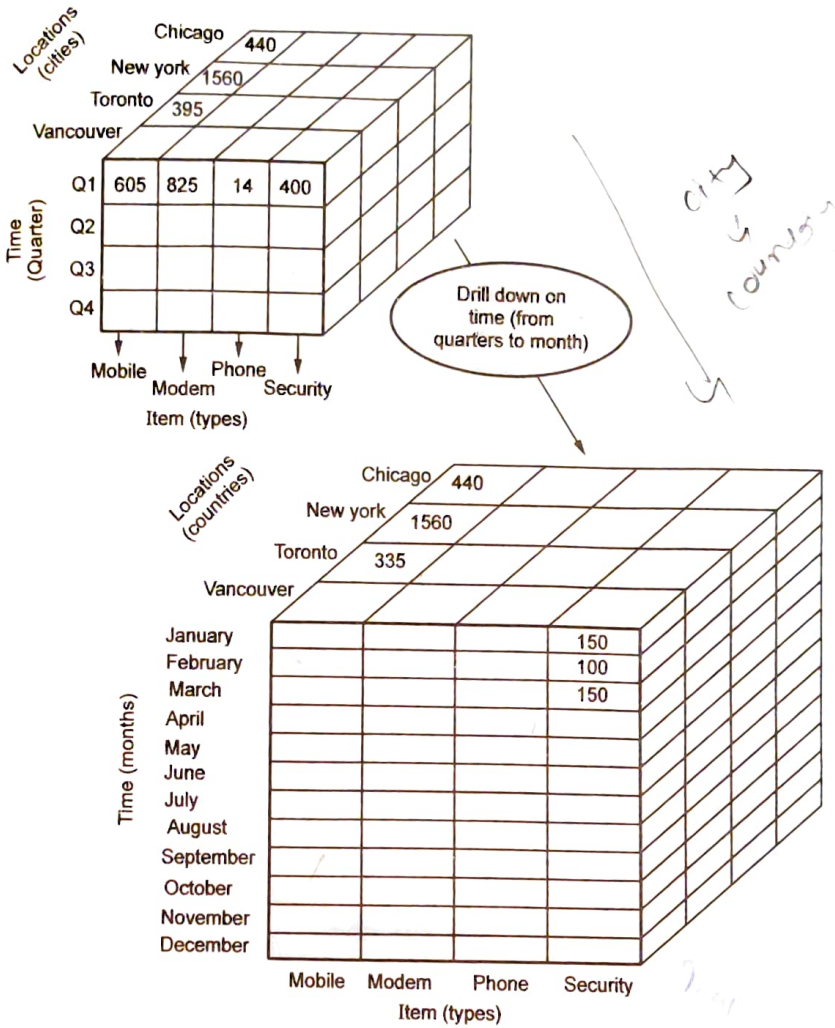


Fig. 2.5.3 Roll drill-down

2.5.3 Slice

The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Fig. 2.5.4 shows how slice works.

- Here slice is performed for the dimension "time" using the criterion time = "Q1".
- It will form a new sub-cube by selecting one or more dimensions.

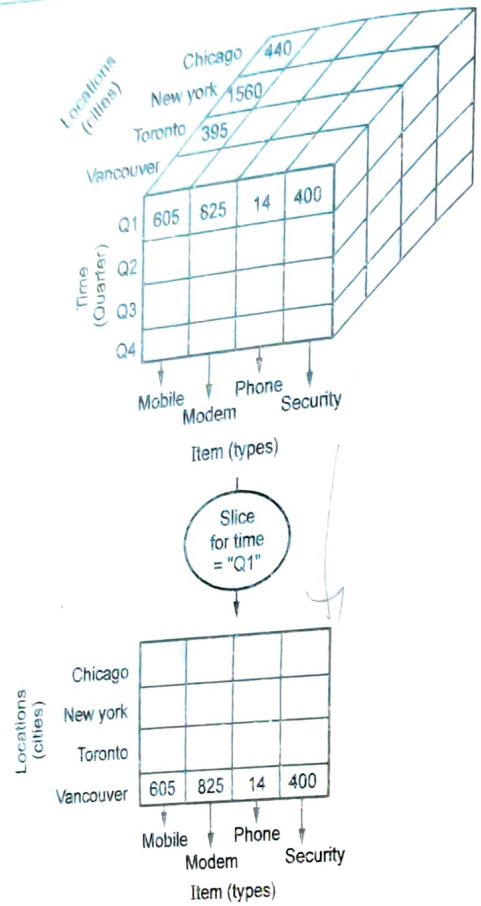


Fig. 2.5.4 Slice

2.5.2 Dice

Dice selects two or more dimensions from a given cube and provides a new sub-cube. Fig. 2.5.5 that shows the dice operation.

(See Fig. 2.5.5 on next page.)

The dice operation on the cube based on the following selection criteria involves three dimensions.

- (location = "Toronto" or "Vancouver")
- (time = "Q1" or "Q2")
- (item = "Mobile" or "Modem")

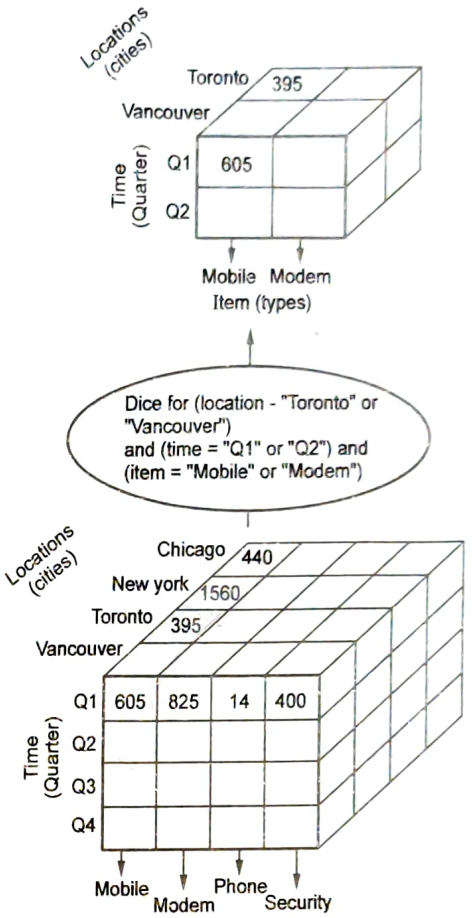


Fig. 2.5.5 Dice

2.5.3 Rotation (pivot)

The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data. Consider the following diagram that shows the pivot operation.

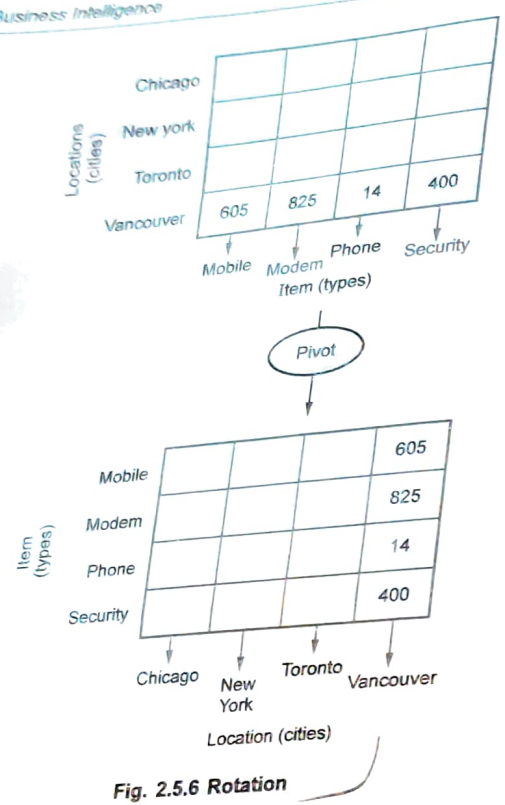


Fig. 2.5.6 Rotation

2.5.4 ROLAP versus MOLAP

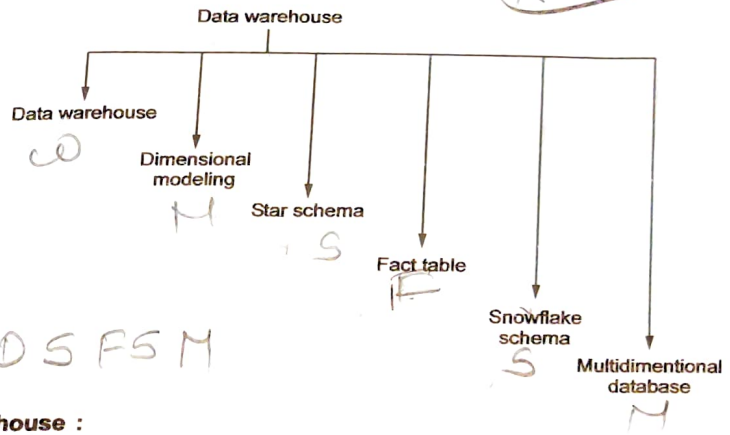
MOLAP (Multidimensional Online Analytical Processing)

The MOLAP storage mode causes the aggregations of the partition and a copy of its source data to be stored in a multidimensional structure in Analysis Services when the partition is processed. This MOLAP structure is highly optimized to maximize query performance. The storage location can be on the computer where the partition is defined or on another computer running Analysis Services. Because a copy of the source data resides in the multidimensional structure, queries can be resolved without accessing the partition's sources data. Query response times can be decreased substantially by using aggregations. The data in the partition's MOLAP structure is only as current as the most recent processing of the partition.

2.5.4.1 ROLAP (Relational Online Analytical Processing)

The ROLAP storage mode causes the aggregations of the partition to be stored in indexed views in the relational database that was specified in the partition's data source. Unlike the MOLAP storage mode, ROLAP does not cause a copy of the source data to be stored in the Analysis Services data folders. Instead, when results cannot be derived from the query cache, the indexed views in the data source are accessed to answer queries. Query response is generally slower with ROLAP storage than with the MOLAP or HOLAP storage modes. Processing time is also typically slower with ROLAP. However, ROLAP enables users to view data in real time and can save storage space when you are working with large datasets that are infrequently queried, such as purely hostorical data.

2.5.5 The Building Blocks : Defining Features



a) Data warehouse :

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other source. It separates analysis workload from transaction workload and enables an organization to consolidate data from several source.

- **Subject oriented :** Data in an organization is organized in major objects or business process. The common example of subject oriented data are customer product, vender and sale transaction.
- **Integrated :** Integration is closely related to subject orientation data warehouse must put data from disparate source into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure, when they achieve this goal, they are said to be integrated.

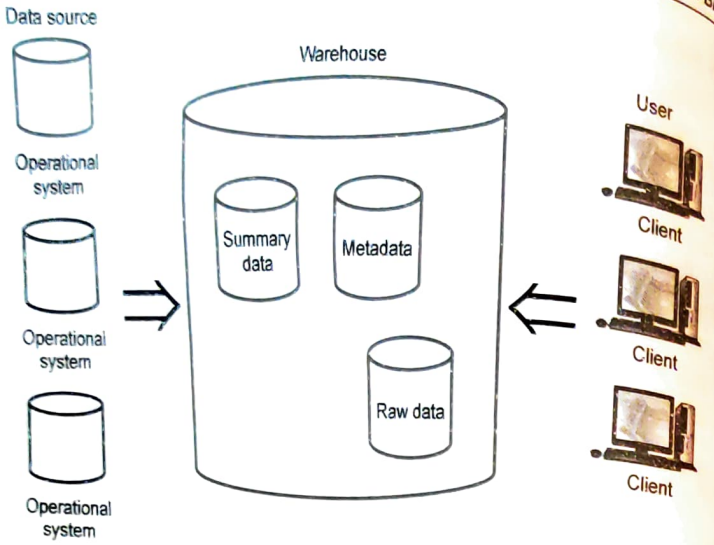


Fig. 2.5.7 Data warehouse

- **Non-volatile** : A data warehouse is always a physically separate store of da transformed from the application data found in operational environment.
- **Time varient** : Data in data warehouse associates with time. The time can be single moment of span of time.

b) Dimensional modeling :

Dimensional modeling is a database design technique to support business users query data in data warehouse. The dimensional modeling is developed to be oriente around query performance and ease of use. It is important to note that the dimensiona modeling is not necessary depends on relational database. The dimensional modelin deling approach is at logical level, which can be applied for any physical forms such a relational and multidimensional database.

In dimensional modeling, there are two important concepts.

i) Facts are also known as business measurement facts are normally numeric value which could be aggregated.

ii) Dimensions are called context : Dimension are business descriptors which specif the facts.

Process of Dimensional Modeling

The following four-step proces is commonly used in dimensional modeling design.

i) Select the business

ii) Declare the grain

iii) Identify the dimensions

iv) Identify the fact.

c) Star schema :

Star schema is a dimensional design for a relational database often used in a data warehouse system. There is a fact table at the center of the schema surrounding by a number of dimension tables therefore the name star schema comes from appearance.

Star Schema Example :

- At the center of the schema we have a fact table FACT-SALES. The primary key of the fact table contains three surrogate keys associated with dimension table.
- Surrounding the fact table is number of dimension table DIM-DATE, DIM-STORE, DIM-PRODUCT.

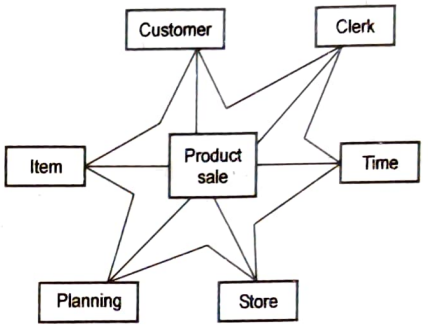


Fig. 2.5.8 Star schema

Star Schema

- The more dimension table you add to the star schema, the more reporting possibilities it provides.
- The capability to study facts only depend on level of detail that the fact table stores.
- Star schema can help business analysis.

d) Fact table :

A fact table is usually in dimensional model in data warehouse design. It is often found at the center of a star schema surrounded by dimension table. Fact table consist of facts of a particular business process sale volume by month by product, fact are also known as measurement.

e) Snowflake schema :

The snowflake schema is a variation of star schema structure normalized use in data warehouse.

The snowflake schema is more complex than the star schema because use of outrigger table, dimension table hierarchies which describe the dimension are normalized.

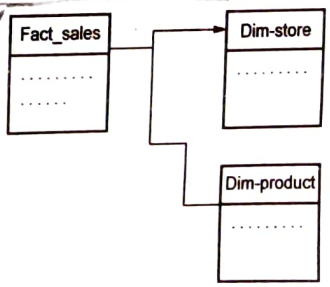


Fig. 2.5.9 Fact table

Example :

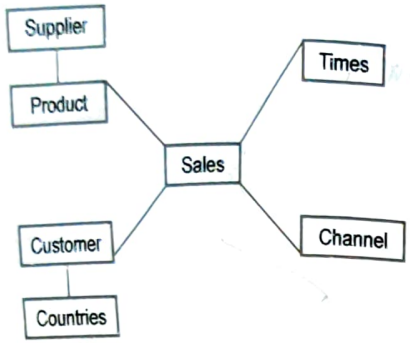


Fig. 2.5.10 Showflake schema

f) Multidimensional database : Multidimensional database is a database has be constructed with the multiple dimension of pre-filled in hyper dimensional cubes of da rather than the traditional two dimensional tables of relation database are as followed.

1) Intimately related

2) Stored viewed and analyzed from different perspectives

Example :

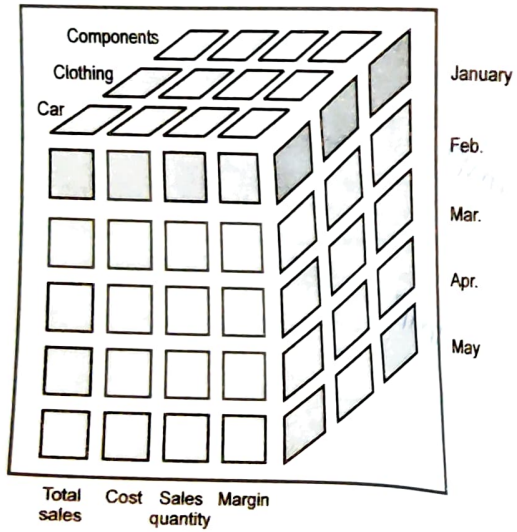


Fig. 2.5.11 Measurement

Review Questions

1. *Define data mining.*
2. *What is data warehouse ?*
3. *Explain the data warehouse architecture.*
4. *Why need for data warehouse ?*
5. *Difference between OLTP and OLAP system.*
6. *Explain star, snowflake, fact constellation.*
7. *What is cubes.*
8. *What is business intelligence ? Explain.*

❑❑❑

3 Introduction to Data Mining (DM)

Syllabus

Motivation for data mining - Data mining-definition and functionalities - Classification of DM systems - DM task primitives - Integration of a data mining system with a database or a data warehouse - Issues in DM - KDD process.

Contents

3.1 Introduction

- Data mining is concerned with solving problem by analyzing existing data.
- "Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from huge amount of data".
- Alternative name : Knowledge Discovery in Database (KDD).
- Data mining is the process of identifying valid, novel, potentially useful and ultimately understandable pattern in the data .

3.1.1 Major Issues in Data Mining

- Mining methodology and user interaction
  - Mining different kinds of knowledge in database.
  - Interactive mining of knowledge at multiple levels of abstraction.
  - Incorporation of background knowledge.
  - Data mining query language and Ad-hoc data mining.
  - Expression and visualization of data mining results.
  - Handling noise and incomplete data.
  - Pattern evaluation : The interesting problem.
- Performance and scalabilty.
  - Efficiency and scalabilty of data mining algorithms.
  - Parallel, distributed and incremental mining problem.
- Issue relating to the diversity of data type
  - Handling relational and complex types of data mining information from heterogeneous database and global information system(www).
- Issue related to application and social impacts
  - Application of discovered knowledge Domain specific data mining tools.
  - Intelligent query processing
  - Process control and decision making.
  - Integration of the discovered knowledge with existing knowledge : knowledge fusion problem protection of data security, integrity and privacy.

3.2 Motivation for Data Mining

Data mining systems depend on database to supply the raw input and this raise problems. Such as that database tend to be dynamic, incomplete, noisy and large other problems arise as a result of the inadequacy and irrelevance of the information stored. The difficulties in data mining can be categorized as :

i) Limited information
ii) Noise or missing data
iii) User interaction and prior knowledge.
iv) Uncertainty
v) Size, updata and irrelevant fields.

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years. Due to wide availability of huge amount of data and imminent need for turning such data into userful information and knowledge. The information and knowledge gained can be used for application ranging from market analysis, fraud detection and customer retention to production control and science exploration.

Data mining can be viewed as a result of the natural evolution of information technology. The database system industry has witnessed and evolutionary path in the development of the following functionalities.

Data collection and database creation, data management and advanced data analysis.

Since the 1960, database and information technology has been evolving systematically from primitive file processing systems to sophisticated and powerful database system. The research and development in database system since the 1970's has progressed from early hierarchical and network database systems to the development of relational database system, data modeling tools and indexing and accessing methods, database technology. Since the mid 1980 has been characterized by the popular adoption of relation technology and an upsure of research and development activities on new and powerful database system. The steady and amazing progress of computer hardware technology in the past three decades has led to large supplies of powerful and affordable computer, data collection equipment and storage media, this technology provides a great boost to the database and information repositories available for transation management, information retrieval and data analysis. Data an now be stored in many different kinds of database and information repositories.

The abundance of data coupled with the need for powerful data analysis tools, has been described as a data rich but information poor situation, the fast growing tremendous amount of data collected and stored in large and numerous data repositories has for exceeded our human ability for comprehension without powerful tools. Data collected in large data repositories become "data tombs" data archies that are seldom visited datamining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases and scientific and medical research. The widening gap between data and information call for a systematic

development of data mining tools, that will turn data tombs into "golden nuggets" of knowledge.

3.3 Data Mining Definition and Functionalities

- Data mining is concerened with solving problem by analying existing data.
- "Extraction of interesting (non-trival implicit, previously unknown and potentially usefull information or patterns from large amount of data".
- Alterenative names : Knowledge Discovery in Database (KDD)
- Data mining is process of identifying valid, novel potentially useful and useful and ultimately understandable pattern in the data.

Data mining provides the link between the two types : a) Data mining software analysis relationship b) Patterns in stored transition databased on open-ended user queries.

Several types of analysis software are available; statical, dynamic, machine learning and neural network, fuzzy logic, generally any of four type of relationship are sought following architecture of data mining.

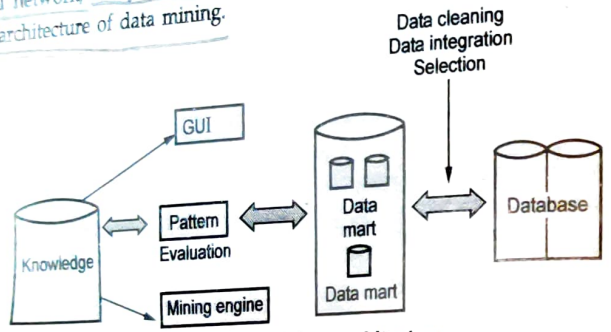


Fig. 3.3.1 Data mining architecture

Database, data warehouse information :

This is one or a set of database, data warehouses. Spreadsheets or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the database.

Database or data warehouse server :

The database or data werehouse server is responsible for fetching the relevant data, based on the user's data mining request.

Knowledge base :

This is the domain knowledge that is used to guide the search or evaluate. The interestingness of resulting patterns such knowledge can include concept hierarchies

used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to access a patterns interestingness based on its unexpectedness, may also be include; other example of domain knowledge are additional interestingness constraints or thresholds and metadata.

Data mining engine :

This is essential to the data mining system and ideally consist of a set of functional modules for tasks as characterization, association analysis, classification and evolution and deviation analysis.

Pattern evaluation module :

This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search towards interesting patterns. It may the access interestingness thresholds stored in the knowledge base. The pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used for efficient data mining. It is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process so as to confine the search to only the interesting patterns.

Graphical user interface :

This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search and performing exploratory data mining based on the intermediate data mining results. This component allows the user to browse database and data warehouse schemes or data structures, evaluate mined patterns and visualized the pattern in different forms.

3.4 Classification of DM (Datamining) Systems

Classification problem aim to identify the characteristics that indicate the group to which each case belongs. The pattern can be used both to understand the existing data and to predict how new instances will behave. For example, you may want to predict whether individuals can be classified as likely to respond to a direct mail solicitation. vulnerable to switching over to a competing long distance phone service or a good candidate for a surgical procedure. Data mining creates classification model by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from an historical database, such as people who have already undergone a particular medical treatment or moved to a new long distance service. They may come from an experiment in which a sample of the entire database is tested in the real world and the result used to create a classifier. For example : a sample of a mailing list would be sent an offer and the result of the mailing used to develop a

classification model to be applied to the entire database. Sometimes an expert classifies a sample of the database and this classification is then used to create the model which will be applied to the entire database.

Classification belongs to the category of supervised learning, distinguished from unsupervised learning. In supervised learning, the training data consist of pairs of input data (typically vectors) and desired output, while in unsupervised learning there is no a priori output.

Classification has various application, such as learning from a patient database to diagnose a disease based on the symptoms of a patient, analyzing credit card transactions to identify fraudulent transations, automatic recognition of letters or digits based on handwriting samples and distinguishing highly active compounds from inactive ones based on the structures of compounds for drug discovery. Classification has been studies in statics and machine learning. In statics classification is also referred to as discrimination.

Classifying an example based on distances in the space (the k-nearest neighbour method) and constructing a classification tree that classifies example based on tests on one or more predictor variable (classification tree analysis). In the field of machine learning, attention has more focused on generating classification expressions that tree learning, which learns the same tree structure as classification tree but uses different criteria during the learning process. The technique was developed in parallel with the classification tree analysis in statics. Other machine learning techniques include classification rule learning neural network, Bayesian classification instance-based learning, genetic algorithms the rough set approach and support vector machines. These techniques mimics human reasoning in different aspects to provide insight into the learning process. The data mining community inherits the classification techniques developed in statics and machine learning and applies them to various real world problem most statical and machine learning algorithms are memory based, in which the whole training data set is loaded into the main memory before learning starts. In data mining, much effort has been spent on scaling up the classification algorithms to deal with large data sets. There is also a new classification technique, called association based classification, association rule learning.

3.4.1 Classification vs Predication

Classification : Predicts categorical class labels (discrete or normal) classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data.

Prediction : Models continuous valued function.

Predicts unknown or missing values.

Typical application : Credit approval

Target marketing

Medical diagnosis

Fraud detection

3.4.2 Classification

A two-step process :

I) Model construction : Describing a set of predetermined classes.

a) Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute.

b) The set of tuple used for model construction is training set.

c) The model is represented as classification rules, decision trees or mathematical formulae.

II) Model usage : For classifying future unknown object estimate accuracy of the model.

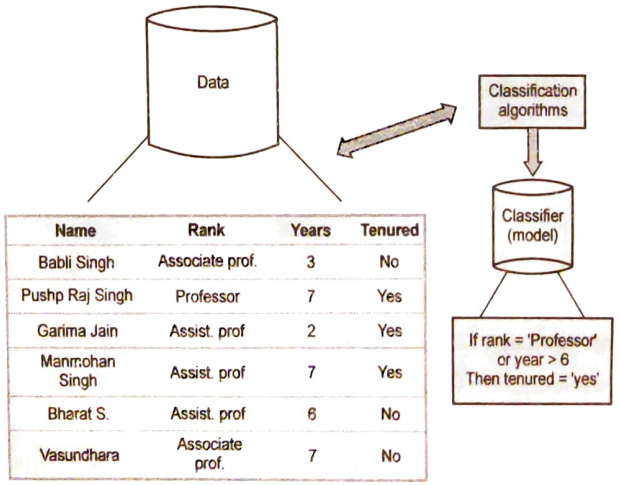
a) The known label of test sample is compared with the classified result from the model.

b) Accuracy rate is the percentage of test set sample that are correctly classified by the model.

c) Test set is independent of training set, otherwise over-fitting will occur.

d) If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known.

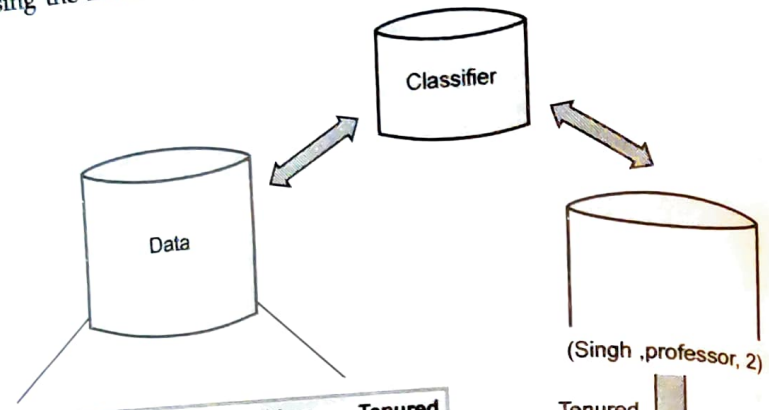
Process (1) : Model construction



Name	Rank	Years	Tenured
Babli Singh	Associate prof.	3	No
Pushp Raj Singh	Professor	7	Yes
Garima Jain	Assist. prof	2	Yes
Manmohan Singh	Assist. prof	7	Yes
Bharat S.	Assist. prof	6	No
Vasundhara	Associate prof.	7	No

Fig. 3.4.1 Model construction process

Process (2) : Using the model in prediction.



Name	Rank	Year	Tenured
Babli Singh	Associate prof.	2	No
Anwati S.	Professor	7	No
Rahul	Assist. prof.	5	Yes
Satish	Assist. prof.	7	Yes

Fig. 3.4.2 Using the model in prediction

3.4.1 Supervised vs Unsupervised Learning

I) Supervised learning (Classification)

1. Supervision : The training data (observation, measurements) are accompan by labels indicating the class of the observations.
2. New data is classified based on the training set.

II) Unsupervised learning (Clustering)

1. The class labels of training data is unknown.
2. Given a set of measurement, observation, with the aim of establishing existence of classes or cluster in the data.

3.4.2 Data Preparation

1. Data cleaning : Preprocess data in order to reduce noise and handle missi values.
2. Relevance analysis : Remove the irrelevant or redundant attributes.
3. Data transformation : Generalize and/or normalize data.

3.5 DM (Data Mining) Task Primitives

A data mining task can be specified in the form of a data mining query which is input to datamining system. A data mining query is defined in terms of data mining task primitives. These primitives allows the user to interactively communicate with data mining system during discovery in order to direct the mining process or examine the findings from different angles or depths. The data mining primitive specify the following.

- **The set of task - relevant data to be mined :**
  This specifies the portion of the database or the set of data in which the user is interested. This includes data base attributes or data warehouse dimensions of interest.
- **The kind of knowledge to be mined :**
  This specifies the data mining function to be performed. Such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis or evolution analysis.
- **The background knowledge to be used in the discovery process :**
  The knowledge about the domain to be mined is useful for guiding the knowledge discovery process and for evaluating the patterns found. Concept hierarchy are a popular form of background knowledge, which allow data to be mined at multiple levels of abstraction.
- **The interesting measures and thresholds for pattern evaluation :**
  They may be used to guide the mining process, or after discovery to evaluate the discovered pattern. Different kinds of knowledge may have different interesting measure, for example, interestingness measure for association rule include support and confidence.
- **The expected representation for visualization the discovered patterns :**
  This refers to the form in which discovered patterns are to be displayed, which may include rules, charts, graphs, decision trees and cubes.

A datamining query language can be designed to incorporate these primitive, allowing user to flexibily interact with data mining system. Having a data mining query language provides a foundation on which user, friendly graphical interfaces can be built.

Task-relevant Data

Database or data warehouse name

Database tables or data warehouse cubes

Condition for data selection

Relevant attributes

Data grouping crieteria

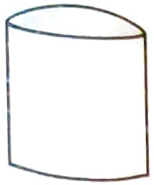


Fig. 3.5.1

Knowledge type to be mined

Characterization

Discrimination

Association

Classification

Clustering

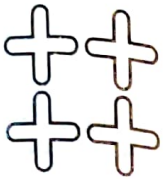


Fig. 3.5.2

Background knowledge

Concept hierarchies

User beliefs about relationships in the data

Fig. 3.5.3

Pattern interestingness measures

Simplicity

Certainty

Utility

Novelty

Fig. 3.5.4

Visualization of discovered patterns

Rules, tables, reports, charts, graph, trees and cubes

Drill-down and roll up

Fig. 3.5.5

3.6 Integration of a Data Mining System with a Data Base or Data Warehouse

A fundamental concept of data warehouse is the distinction between data an information in system. **Data** are only facts, numbers or text can be processed by computer. **Information** is an integrated collection of facts and is used as the basis f decision making purpose. Data mining is described as a purpose of discover extracting interesting knowledge from large amount of data stored in multiple da sources as file systems. Data mining system work as a stand-alone system in

application program of data warehouse system with which it has to communicate this simple schemes is called coupling data mining system retrieves data from a particular and different source such as file system can be following several different-diffrent coupling.

No-coupling : Data mining system does not utilize any functionality of a database system. A no-coupling data mining system retrieves data from a individual data sources such a file system. Process data using major data mining any algorithm and stores result into file system this type system is also called **No-coupling**.

Such a system, has simple drawback is as follows :

i. Data warehouse that is already very efficient in organizing data for (cleaning, indexed, integrated) storing, accessing and retrieving data.

ii. It is not possible to data can be retrieve always single source.

iii. No-coupling architecture is very poor for data mining system.

- **Loose - coupling :**
  Loose coupling means that a data mining use database or data warehousing for data retrieval from a data repository managed by these systems.

  Loose-coupling mainly for memory-based because data mining system that does not explore data structure and query optimization method provide by data base system.

- **Semitight coupling :**
  Semi-tight coupling data mining system, beside linking to database system. Performs some data mining tasks (including sorting, indexing, aggregation, efficient implementation) primitives can be provided in data base system for better then performance.

- **Tight - coupling :**
  Tight-coupling data mining subsystem is treated as an functional component of data mining system using integration. Other feature of database or data warehouse query analysis, data structure, indexing schemas and query processing perform data mining task.

  Data mining system provides scalability and high performance.

3.7 Issues in Data Mining

A data mining initiatives continue to evolve, there are several issues congress may decide to consider related to implementation and oversight. These include, but are not limited to data quality, interoperability, mission creep and privacy. As with other aspects of data mining, while technological capabilities are important.

i) **Data quality** : Data quality is a multifaceted issue that represents one of th biggest challenges for data mining. Data quality refers to the accuracy an completeness of the data. Data quality can also be affected by the structure an consistency of the data being analyzed. The presence of duplicate records the la of data standards, the timeliness of updates and human error can significant impact the effectiveness of the more complex data mining techniques, which ar sensitive to subtle differences that may exist in the data to improve data qualit It is sometimes necessary to "clean" the data, which can involve the removal duplicate records, normalizing the values used to represent information in th database (e.g. ensuring that 'no' is represented as a 0 throughout the database an not sometimes as a 0, sometimes as a N), accounting for missing data point removing unneeded data fields, identifying anomalous data points an standardizing data formats.

ii) Interoperability : Related to data quality is the issue of interoperability o different database and datamining software. Interoperability refers to the ability o a computer system and/or data to work with other system or data using commo standards or process. Interoperability is a critical part of the larger efforts t improve interagency. Collaboration and information sharing through e-governmen and homeland security initiatives. For data mining interoperability of databas and software is important to enable the search and analysis of multiple databas simultaneously. Ensure the compatibility of data mining activities of differen agencies. Data mining project that are trying to take advantage of existing legac database or that are initiating first time collaborative efforts with other agencies o levels of government may experience interoperability problems as agencies mov forward with the creation of new database and information sharing efforts, the will need to to address interoperability issues during their planning stages t better ensure the effectiveness of their data mining projects.

iii) **Mission creep** : Mission creep is one of the leading risks of datamining cited b civil libertarian and represents how control over one's information can be tenuous proposition. This can occur regardless of whether the data was provide voluntary by the individual or was collected through other means. Efforts to figh terrorism can, at times take on an acute sense of urgency. This urgency can creat pressure on both data holders and officials who access the data to leave a available resource unused may appear to some as being negligent. Data holde may feel obligated to make any information available that could be used t prevent a future attack or track a known terrorist, government official responsibl for ensuring the safety of others may be pressured to use and/or combine existin database to identify potential threats. Unlike physical searches or the detention o

individuals, accessing information for purposes other than originally intended may appear to be a victimless or harmless exercise.

Such information use can lead to unintended outcomes and produce misleading results. One of primary reasons for misleading results is inaccurate data.

iv) **Privacy** : As additional information sharing and data mining initiatives have been announced, increased attention has focused on the implications for privacy, as well as concerns about the potential for data mining application to be expanded beyond their original purpose.

For example : Some experts suggest that antiterrorism data mining application might also be useful for combating other types of crime as well so far there has been little consensus about how data mining should be carried out, with several competing points of view being debated. Some observers contend that tradeoffs may need to be made regarding privacy to ensure security.

As data mining efforts move forward, congress may consider a variety of questions including, the degree to which government agencies should use and mix commercial data with government data, whether data sources are being used for purposes other than those for which they were originally designed and the possible application of the privacy act to these initiatives.

v) **Performance issues** : Many artificial intelligence and statical methods exist for data analysis and interpretation. These methods were often not designed for the very large data sets data mining is dealing with today. This raise the issues of scalability and efficiency of the data mining methods when processing considerably large data. Algorithms with exponential and even medium order polynomial complexity cannot be of practical use for data mining. Sampling can be used for mining instead of the whole data set. Concerns such as completeness and choice of samples may arise.

vi) **Time series** : Time series forecasting predicts unknown future values based on a time varying series of predictors it uses known results to guide its predictions. Model must take into account the distinctive properties of time especially the hierarchy of periods (including such varied definitions as the five or seven day work week, the thirteen "month" year), seasonality, calender effects such as holidays, data arithmetic and special considerations such as how much of the past is relevant..

vii) **Pattern evaluation** : A data mining system can uncover thousands of pattern many pattern discovered may be uninteresting to the given user.

3.8 KDD Process-Various Model and their Significance

With the enormous amount of data stored in files database and other repositories, it is increasingly important. If not necessary to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Datamining also popularly known as Knowledge Discovery in Database (KDD), to the nontrivial extraction of implicit. Previously unknown and potentially useful information from data in database. While data mining and knowledge discovery in Database (KDD) are frequently treated as synonyms datamining is actually part of the knowledge discovery process The Fig. 3.8.1 shows iterative knowledge discovery process.

The knowledge discovery in database process comprises of a few steps learning from raw data collection to some from of new knowledge. The iterative process consist of the following steps.

- **Data cleaning** : Also known as data cleaning, it is a phase in which noise data and irrelevant data are removed from the collection.
- **Data integration** : At this stage, multiple data source, often heterogeneous, may be combined in a common source.
- **Data selection** : At this stage, the data relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation** : Also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining** : It is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- **Pattern evaluation** : In this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation** : Is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

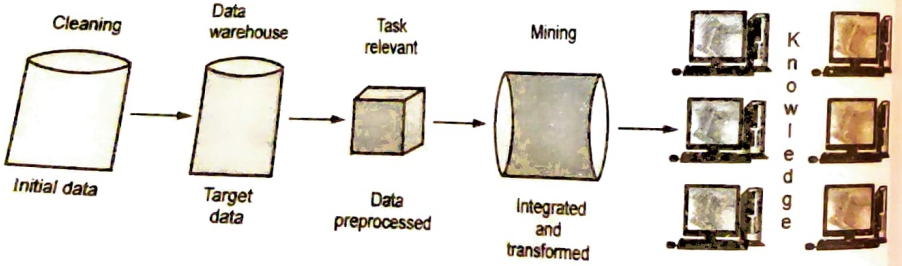


Fig. 3.8.1 KDD various model

It is common to combine some of these steps together, for instance data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse. Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, as for the case of data warehouse, the selection is done on transformed data.

The KDD is an iterative process once the discovered knowledge is presented to the user, the evaluation measure can be enhanced, the mining can be further refined. New data can be selected or further transformed or new data sources can be integrated, in order to get different. More appropriate results, data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either sifting through a large amount of material or ingeniously. Probing the material to exactly pinpoint where the values reside. It is however a misnomer since mining for gold in rocks is usually called "gold mining" and not "rock mining", thus by analogy, data mining should have been called "knowledge mining" instead. Data mining become the accepted customary term and very rapidly a trend that even overshadowed more general terms such as Knowledge Discovery in Database (KDD) that describe a more complete process. Other similar terms referring to data mining are : data dredging, knowledge extracting and pattern discovery.

3.8.1 Application Challenges for KDD

- Overfitting
- High dimensionally
- Larger database
- Understandability of patterns
- Mission on noisy data.

3.8.2 AI in KDD

- Natural language processing
  - Queries
  - Interface
  - Free from text mining.
- Intelligent agents
  - Data collection
  - Remote operation.
- Knowledge representation
  - Ontologies

- Based on human knowledge.
- Uncertainty in AI
  - Interface
  - Reasoning.

3.8.3 Current Application on KDD

- **Science :** Used to aid astronomers by classifying faint sky object.
- **Marketing :** Used customer group identification and forecasting.
- **Fraud detection :** Credit card fraud detection.
- **Data cleaning :** Use by India some state to locate and remove duplicate welfare claims.

Review Questions

1. *Explain motivation for data mining.*
2. *What is major issues in data mining ?*
3. *What is classification ?*
4. *Discuss various data mining issues.*
5. *Explain KDD process.*

10, 2, 19, 18, 20, 18, 25, 28, 22

Sorting :- 2, 10, 18, 18, 19, 20, 22, 25, 28

Bin1 Bin2 Bin3

10 19 25

x =

dummy value - 10, 10, 19, 10, 19, 19, 25, 25, 25

4 Data Pre-processing

Syllabus

Why to pre-process data ? - Data cleaning : Missing values, Noisy data - Data integration and transformation - Data reduction : Data cube aggregation, Dimensionality reduction - Data compression - Numerosity reduction - Data mining primitives - Languages and system architectures : Task relevant data - Kind of knowledge to be mined - Discretization and concept hierarchy.

Contents

6, 9, 12, 13, 15, 25, 50, 70, 72, 92, 204, 232

6, 9, 12 | 13, 15, 25 | 50, 70, 72 | 92, 204, 232

9 19 61 162

dv :- 6, 9, 12, 13, 15, 25, 50, 70, 72, 92, 204, 232

9 9 9 19 19 19 61 61 61 162 162 162

4.1 Data Pre-Processing

- Analyzing data that has not been actually represent such problem it represent only such misleading results.
- Data pre-processing there is much irrelevent and repetance such information be present.
- Data pre-processing is a represent only training set.
- pre-processing has include cleaning, normalization, transformation, feature extraction.
- For quality of mining result needed.

4.1.1 Why Data-Pre-Processing can be Required ?

- Why data pre-processing for quality of mining result quality of data is needed.
- Preparation and transformation of initial level in data set.
- Raw data is highly and important step for successful data mining.
- Aggregate information : Useful to be obtain aggregate information.
- Enhancing mining : It can be reduce the number of data set to enhance improvement performance of mining process.
- Noise : It is very important to use some techniques to reduce noise data.
- Improve quality of data : To improve accuracy, efficiency and knowledge discovery process can be subsequent of data mining process.

4.1.2 Data Pre-Processing Tasks

- Data cleaning :
  - Missing data may need to be inferred.
  - Identify outliers and noise data.
  - Inconsistent data.
- Data transformation :
  - Consolidate data into forms suitable for data mining.
  - Important to save normalization parameter.
  - It to be handle or represent decimal point for all value.
- Data parsing and standardization :
  - Based on data to be links.
  - Use to well define attribute.
- Data integration :
  - Using multiple database.

4.1.3 Data Pre-Processing Quality

- Accuracy
- Completeness
- Consistency
- Accessibility
- Timeless

4.2 Various Forms of Data Pre-Processing

i) Data cleaning
ii) Data integration
iii) Data transformation
iv) Data reduction
v) Discretization and concept hierarchy generation.

4.3 Data Cleaning

- Data to be completed in incomplete or missing value like noise, inconsistent, identifying outlier.
- Data cleaning are common place properties of large, real word database and data warehouse.

Missing or incompleted value : Incompleted/missing value can be filled by different-different value of attribute like such technique can followed.

i) Avoid tuple
ii) Self adjustment by incompleted value
iii) Global technique can be use
iv) Most suitable value can be fill by the missing value.

- Noise data :
  - Dynamically error in a measurement.
  - Remove noise data through banning of numeric data, clustering, regression and self human inspection.
- Inconsistent Data :
  - Data inconsistent to be error remove by self human using external references.
  - Important to have been due to handle data entry verification.

Why is Data Dirty ?

- Human/hardware or software problem.
- Redundance data in record also need to be cleaning.
- Computer error at entry time by human.
- Multipurpose field.
- Error in data transmission.

4.3.1 Data Cleaning Task

- Fill in missing value
- Unified data format
- Converting nominal to numeric
- Correct inconsistent data
- Identify outlier and smooth out noisy data.
- Data acquision and meta data.

4.3.2 Missing Values

Example :

i) Average on gene(SPOT.Prb Cell) and/or condition replicates.

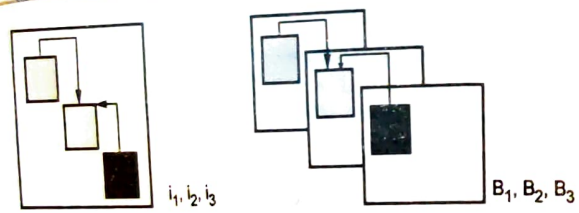


Fig. 4.3.1 Missing values (1)

ii) Average condition : Profile missing entries.

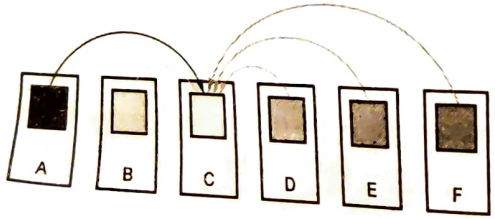


Fig. 4.3.2 Missing values (2)

iii) For time course, interpolation of nearby value in the profile.

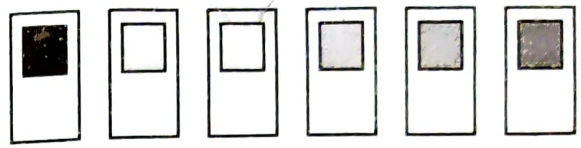


Fig. 4.3.3 Missing values (3)

iv) None of these solution exploits relationship among profile value.

4.3.2.1 Missing Data

- Data is not present always
  - Many tuples have no recorded value for several attribute.
- Missing data it may be due to
  - Data not entered due to misunderstanding
  - Not register history
  - Equipment Malfunction
  - Inconsistent with other recorded data and thus deleted
- Missing data may need to be inferred.

4.3.2.2 Handle of Missing Data

- **Ignore the tuple :** Usually done when class label is missing (assuming the task in classification) not effective when the percentage of missing values per attribute varies considerably.
- **Fill in the missing value by human :** Tedious but it is not feasible for dynamic and large data set.
- **Use global constant to fill in the missing value :** New global constant value "unknown", as a new class.
- **Impuation :** Use the attribute mean to fill in the missing value smarter.

4.4 Noisy Data

- Basically noisy is a variance in a measured variable.
- Missing attribute values may be due to
  - Data entry problem
  - Technology limitation
  - Data transmission problems
  - Inconsistency in naming convention.

- Data problem which requires data cleaning
  - Inconsistent data
  - Incomplete record
  - Duplicate data

4.4.1 Handle Noisy Data

4.4.1.1 Binning Method

- First sort data and partition.
- Then one can smooth by bin mean, smooth by bin median, and smooth by b boundaries.

Method

- Equal-width partition on distance basics
  i) Divide range into N equal size of interval.
  ii) L and H is the lowest and highest value of attribute according width equal size of interval is W = (H – L)/N.
  iii) Most straight forward.
  iv) Outliers may dominate presentation.
  v) Skewed data is do not handle properly.
- Equal-depth partitioning on frequency basic.
  i) Divide range in equal N interval, (each containing similar number samples).
  ii) Using tricky managing categorical attributes.
  iii) Data scaling

Binning Method Example

Example 1 :

Data : 0, 4, 12, 16, 16, 18, 24, 26, 28

A) Equal width :

W = (Max – Min) N or

W = (H – L) N

- Bin 1 : 0, 4 [–, 10]
- Bin 2 : 12, 16, 16, 18 [10, 20]
- Bin 3 : 24, 26, 28 [20, +]

B) Equal frequency :

- Bin 1 : 0, 4, 12 [–, 14]
- Bin 2 : 16, 16, 18 [14, 21]
- Bin 3 : 24, 26, 28 [21, +]

4.4.2 Cluster Analysis

- Outliers may be detected by clustering, where similar values are organized into group or cluster.
- Value that fall outside of the set of cluster may be considered outliers.
- Clustering result are used
  - As a stand-alone tool to get insight into data distribution.
    - Visualization of cluster may unveli important information.
  - As a pre-processing step for other algorithm.
    - Efficient indexing often relies on clustering.
- Each cluster centroid is marked with a '+' represent the average of point in space for that cluster.

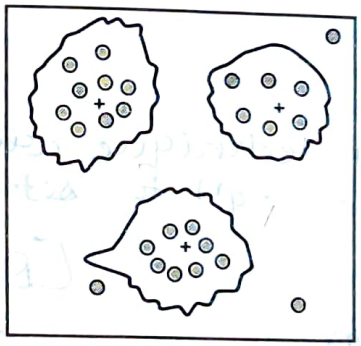


Fig. 4.4.1 Cluster analysis

Resulting graph :

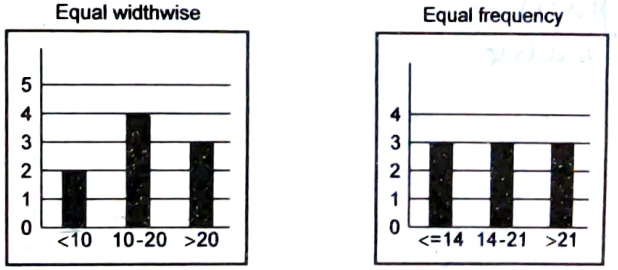


Fig. 4.4.2 Cluster analysis

Example 2 :

Data : 4, 8, 15, 21, 21, 24, 25, 28, 34

1. Partition into equidepth bins :
   - Bin 1 : 4, 8, 15
   - Bin 2 : 21, 21, 24
   - Bin 3 : 25, 28, 34
2. Smooth by bin means :
   - Bin 1 : 9, 9, 9
   - Bin 2 : 22, 22, 22
   - Bin 3 : 29, 29, 29
3. Smooth by bin boundaries :
   - Bin 1 : 4, 4 15
   - Bin 2 : 21, 21, 24
   - Bin 3 : 25, 25, 34

4.4.1.3 Regression

- Data can be smoothed by fitting the data to a function such as regression.
- Using regression on values of attribute fill missing value.

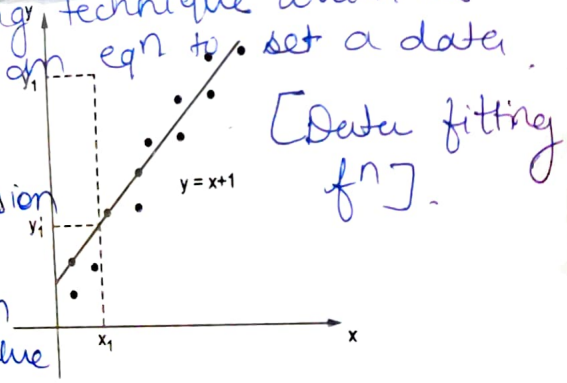


Fig. 4.4.3 Regression

4.4.1.4 Combined Computer and Human Inspection

- Detect suspicious values and check by human.

4.5 Data Integration and Transformation

4.5.1 Data Integration

- Data integration combine data from multiple source into a coherent data store.
- These source may include multiple database, data cube or flat files.

4.5.1.1 Data Integration Issues

- Entity identification problem :
  - Identify real world entities from multiple data source.
- Detecting and resolving data value conflicts.
  - For the same real world entity, attribute, values from different source are different.
- Schema integration and object matching can be tricky.
- Entity identification problem.
  - Equivalent real-world entities from multiple data source be match.
- Redundancy.
  - An attribute may be redundant if it can derived from another attribute.
- Duplication
- Detection and resolution of data value conflicts.

4.5.1.2 Redundancy Causes Problem

Consider like table,

Emp [ENO, ENAME, BASIC, DA, PAY => PAY = BASIC + DA]

EMPLOYEE (ENO, ENAME, BASIC, DA, PF, PAY) ⇒ PAY = BASIC + DA – PF

- If redundant variable are numeric it is better first to normalize them before integrating data from multiple source.

4.5.1.3 Handling Redundancy in Data Integration

- On the basic of redundancy data can occur often when integration of multiple database.
  - Object identification : Object it may have different name in different database.
  - Derivable data : One attribute it may be a "derive d" attribute in other table.

a) Correlation analysis for (numerical data) : Correlation coefficient

$$R_{A,B} = \frac{\sum (A-\bar{A})(B-\bar{B})}{(n-1)\sigma A \sigma B} = \frac{\sum (AB) - n\bar{A}\,\bar{B}}{(n-1)\sigma A \sigma B}$$

where $\bar{A}$ and $\bar{B}$: Respective means of A and B,

$\sigma A, \sigma B$: Respective deviation of A and B

$\Sigma(AB)$: Sum of the AB cross-product.

$R_{AB} > 0$: A and B positive correlated

$R_{AB} = 0$: Independent

$R_{AB} < 0$: Negative correlated

b) Correlation analysis (categorical data)

- $\underset{\text{(chi-Squar)}}{\chi^2}$

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

where : χ^2 : Related variable

Example :

	Play Chess	Not Play Chess	Row Sum
Like Science	250 (90)	200 (360)	450
Not like science	50 (210)	1000 (840)	1050
Sum of (Column)	300	1200	1500

- $\underset{\text{(Chi-squar)}}{\chi^2}$ Calculation : Number in parenthesis are expected counts calculated ba on the data distribution in the categories.

$$\underset{\text{(Chi-squar)}}{\chi^2} = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= 507.93$$ (define correlation group between like science and play-che

4.5.2 Transformation

- Data are transformed or consolidated into the form of befitting or allocate mining.
- Data transformation technique is two type
  - i) Code generation
  - ii) Data mapping maps data element

4.5.2.1 Data Transformation can Involve the Following Activities

- Smoothing : It remove noise from data by using binning and clustering.
- Aggregation : Use for summarization and data cube constrain purpose.

- Generalization : Primitive data are always replaced by high level concept by using of the hierarchies concepts.
- Normalization : When attribute data scaled to fall within a small and specified range, like (– 1.0 to 1.0)
  - Transform V in min-max normalization.
  - Z-score normalization
  - Normalization by using decimal scaling.
- Attribute/feature construction : New attributes as constructed from the given ones new process.

4.5.2.2 Data Transformation Method

i) Transformation V min-max normalization

$$V' = \frac{V - MIN_A}{MAX_A - MIN_A}(New\_Max_A - New\_Min_A) + New\_Min_A$$

Example : Income range 12, 000 to 98, 000 normalization [0.0, 1.0] and then 73, 600 is mapped by

$$= \frac{73600 - 12000}{98000 - 12000}(1.0 - 0.0) + 0$$

$$= 0.716$$

ii) Z-score normalization

$$V' = \frac{V - \mu_A}{\sigma_A}$$

$$V' = \frac{73600 - 54000}{16000} = 1.225$$

where $\mu = 54000$, $\sigma = 16000$

iii) Normalization by using decimal scaling

$$V' = \frac{V}{10^j}$$ when j is smallest integer

4.6 Data Reduction

- Database may store lot of data.
- Very complex data can analysis and mining on large amount of data may be taken by very long time period.
- Reducing technique can be reducing the number of attribute.

4.6.1 Data Reduction Strategies

- Data cube aggregation.
- Dimensionality reduction.
- Numerosity reduction.
- Attribute and concept hierarchy generation.

4.6.2 Data Cube Aggregation

- This operation can be applied to the data in construction of a data cube.
- When cube is generated to the lowest level of data abstraction is called b cuboid.
- When cube highest level of data abstraction is called **apex cuboid**.
- Use small representation which is enough to solve related task.

4.6.3 Dimensionality Reduction

- Data transformation technique can be applied to obtain a compres representation of the base data by using.
  - i) Lossy compression
  - ii) Lossless compression
- Heuristic method for attribute subset selection
  - Step-wise forward selection
  - Step-wise backward elimination
  - Combining forward selection and backward elimination.
  - Decision tree induction

Example : Heuristic method in decision tree induction

When initial attribute set :

$\{A_1, A_2, A_3, A_4, A_5, A_6\}$

∵ There are reduce set : A_1, A_4 and A_6.

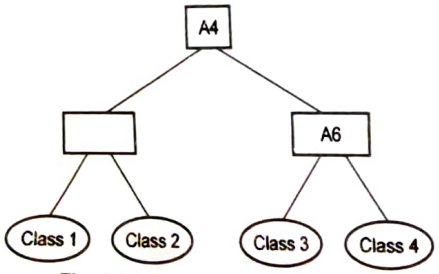


Fig. 4.6.1 Dimensionality reduction

4.6.4 Data Compression

- It is a encoding mechanisms process the amount of data required to represent a given quantity of information.
- Data compression is achieved when redundancies are eliminated.
- Data compression is categories.
  - i) Lossless compression
  - ii) Lossy compression
- Lossless compression
  - Lossless compressed data is reconstructed/extract duplication and the original data is obtain without any loss of information.
  - Lossless compression popular algorithm like LZW (Lempel-Ziv-Welch), RLE (Run-Length-Encoding) Huffman coding, Arithmetic coding, Delta - Encoding.

4.6.5 Lossless Data Compression Algorithm

i) Huffman encoding

ii) Run-Length encoding

iii) Lempel-Ziv-Welch encoding

i) Huffman Encoding : There are two type example

a) Fix length encoding b) Variable length encoding

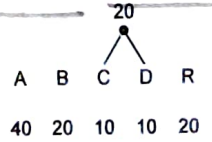
Example :

- Assume that frequency of symbols.

  A = 40, B = 20, C = 10, D = 10, R = 20

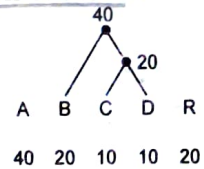
Step 1 :

- Smallest number are 10 and 10 where is C and D, so connected them,



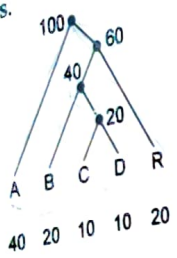
Step 2 : Then smallest values are B, C + D and R, all of which have value 20.

When it is clear that the algorithm does not construct a unique tree, but even if we have decide other possible connection.



Step 3 : Smallest value is R, while A and B + C + D have value 40 then R connects to others.

- Connected the final nodes.



ii) Run-Length-Encoding :

Example :

Original : 17 8 54 0 0 0 97 5 16 0 45 23 0 0 0 0 0 3 67 0 0 0 8 .....

RLE : 17 8 54 0 . 3 97 5 16 0 1 45 23 0 5 3 6 7 0 3 8 .....

) Lempel-Ziv-Welch Encoding :

Example :

Algorithm is working from left to right and has already encoded the left part string will be S = bacaabcde, string E = aabdebb is the data yet to be encoded.

bacaabcdeaabdebb...

4.6.6 Lossy Compression Data

- Lossy compression can be compressed data is reconstructed when the matchi data is obtain and loss of information occur.
- This technique use to original message cannot be recovered exactly as it w before compressed.
- There are lossy compression technique
  - i) Wavelet transforms
    - Discrete wavelet transform
    - Discrete fourier transform
    - Hierarchical pyramid transform algorithm.
  - ii) Principle Component Analysis (PCA)

4.6.6.1 Wavelet Transform

- Discrete wavelet transform : Based on linear signal processing, multi resolutional analysis.
- Compressed matching : Store only for a small compression ratio of the strongest of the wavelet coefficients.
- Discrete fourier transform : Best for define lossy compression.

- Technique or method :
  - If L is length but always power of integer 2.
  - Each transform has two function.
    - i) Smoothing ii) Difference.
  - Applies to pairs data, resulting in two set of length L/2.
  - Applies two function recursively, until reaches the desired length.

Example : For wavelet transformation (See Figure on next page.)

Image Compression Steps

There are three steps of

i) Linear transformation

ii) Quantization

iii) Entropy encoding and decoding

Data Compression Application

i) Image compression

ii) Audio compression

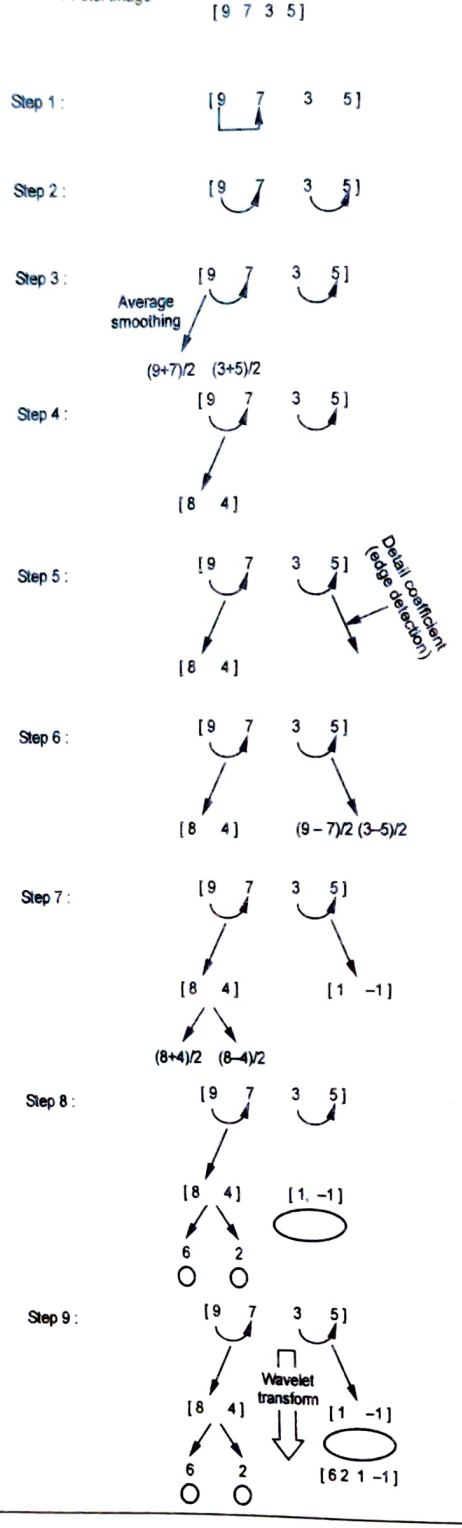
iii) Video compression

iv) String compression

v) General data compression

Example :

$$X = \begin{bmatrix} 100 & 97 & 99 \\ 96 & 90 & 90 \\ 80 & 75 & 60 \\ 75 & 85 & 95 \\ 62 & 40 & 28 \\ 77 & 80 & 78 \\ 92 & 91 & 80 \\ 75 & 85 & 100 \end{bmatrix}$$



- When will be n× p matrix : n (Number of row), p (number of columns), Here 8× 3 matrix.
- A variable - covariance matrix has n = p and is called n-dimensional squar matrix.
- $y = b_1x_1 + b_2x_2 + \ldots b_nx_n$

	x	y
1	– 1.264911064	– 1.78885
2	– 0.632455532	– 0.89443
3	0	0
4	0.632455532	0.894427
5	1.264911064	1.788854
Mean	6.0000	6.0000
Var	1	2

$$r_{x,y} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} = \frac{5.6569}{\sqrt{4\times 8}} = 1$$

4.6.6.2 Principal Component Analysis (PCA)

- Principal of PCA
  - It is a linear projection method can to be reduce the number of all parameters.
  - PCA can based on unsupervised learning.
  - There for work only on numeric data only.
  - The number of dimensions is large but map the data into a space of lower dimensionality
- Properties of PCA
  - Normalize input data : When each attribute fall within the same range.
  - When orthonormal vectors is a K, N is data vector from n-dimensions find $K \leq N$.
  - Vector is a linear combination of the K is a principal component vectors.
  - The principal component are sorted in order of decreasing "significance".
  - When new axes are orthonormal and represent the directions with the maximum variability.

	x	y
x	1	1
y	1	1

Correlation matrix

Covariance matrix

	x	y
x	1	1.414
y	1.414	2

$$\text{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \sqrt{2}$$

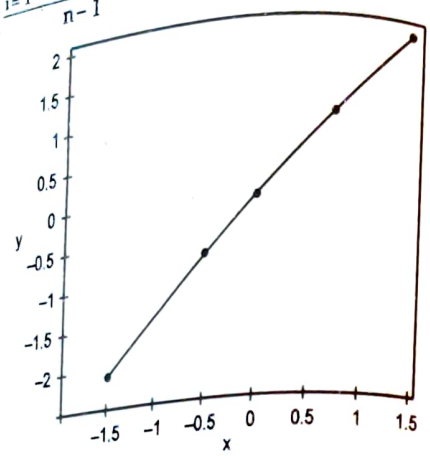


Fig. 4.6.2 DCA

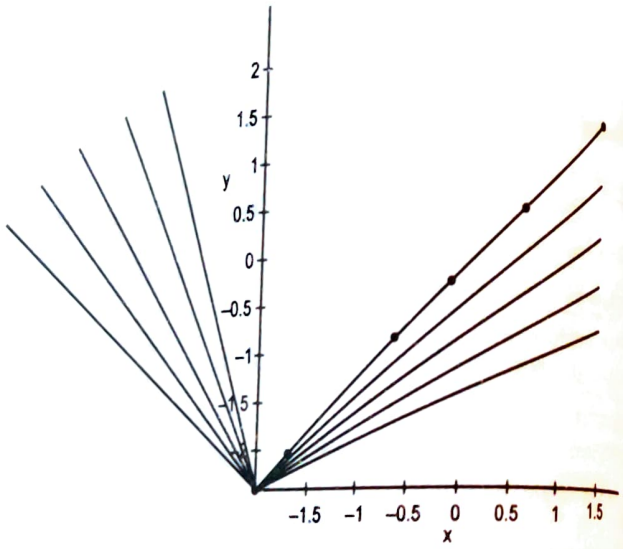


Fig. 4.6.3 PCA

4.6.7 Numerosity Reduction

- Numerosity reduction technique can be reduce the data volume by alternative, smaller forms of data representation.
- Numerosity data reduction use two techniques.
  i) Parametric technique/method

  ii) Non-parametric technique/method

i) Parametric technique :
- Represent instead of the real data.
- It is a parameter estimate model.
- Log-linear model.
- Outliers may also to be stored.

ii) Non-parametric technique :
- Do not representations model
- Family : Histogram, cluster, sampling

4.6.7.1 Data Numerosity Reduction Method

- Regression and log-linear model
- Histograms
- Clustering
- Sampling

Regression and Log-Linear Models for Data Reduction

- Linear regression : Data are modeled to be arrange in straight line.
- Two parameters, α and β is to specify the line and are to be approximated by using the data at hand.
- Least squares criterion to use known different-different value $y_1, y_2 \ldots, x_1, x_2 \ldots$
- Multiple regression
  - Multiple regression is based on linear regression difference only predictors can be use more than two straight line but not display same dimension using pixel.

$$y = b_0 + b_1 x_1 + b_2 x_2 \ldots$$

- **Log-linear model :** This model the association and interaction pattern to all type of categorical variable.
- Former a response is to be identified but no such useful status is to be assigned to any variable in log-linear modeling.
- There are two type of long linear model.
  - Single summary statistics
  - Significance test.
- Probability : $P(a, b, c, d) = \alpha\, ab\beta acx ad\delta bcde$

Example : Find regression equation, also find slope, and intercept

Regression formula : $y = a + bx$

Slope (b) $= \{N\sum xy - (\sum x)(\sum y)/N\sum x^2 - (\sum x)^2\}$

Intercept (a) $= \{\sum y - b(\sum x)\}/N$

x value	y value
60	3.1
61	3.6
62	3.8
63	4
65	4.1

Step 1 : $N = 5$

Step 2 :

x value	y value	x* y	x* x
60	3.1	60 * 31 = 186	60 * 60 = 3600
61	3.6	61 * 3.6 = 219.6	61 * 61 = 3721
62	3.8	62 * 3.8 = 235.6	62 * 62 = 3844
63	4	63 * 4 = 252	63 * 63 = 3969
65	4.1	65 * 4.1 = 266.5	65 * 65 = 4225
Total 311	18.6	1159.7	19359

Step 3 : Using slope formula

Slope (b) $= \{N\sum xy - (\sum x)(\sum y)\}/\{N\sum x^2 - (\sum x)^2\}$

$= ((5) * (1159.7) - (311) * (18.6))/((5) * (19359) - (311)^2)$

$= (5798.5 - 5784.6)/(96795 - 96721)$

$= 13.9/74$

$= 0.19$

Step 4 : Using intercept formula

Intercept (a) $= (\sum y - b\sum x) / N$

$= (18.6 - 0.19(311))/5$

$= (18.6 - 59.09)/5$

$= -40.49/5$

$= -8.098$

Step 5 : Regression equation (y) = a + bx

$= -8.098 + 0.19\ x$

Regression equation (y) = a + bx, variable x = 64

$= -8.098 + 0.19\ (64)$

$= -8.098 + 12.16$

$= 4.06$

4.6.2 Histograms Data Reduction

- Histogram is a representation a two dimensional frequency data occur within represent class interval diagram is called a histograms.
- The light of every bar gives the frequency in the form of a rectangle.
- If the class interval are equal the height of each rectangle is proportional to the corresponding class frequency.
- All way made class interval on x-axis and frequency on y-axis.
- Class interval must be always exclusive.
- The scale for both of the axes need not be same.

Example 1 : Draw a histogram for the following data.

Class interval	Frequency
0 - 5	4
5 - 10	10
10 - 15	18
15 - 20	8
20 - 28	12

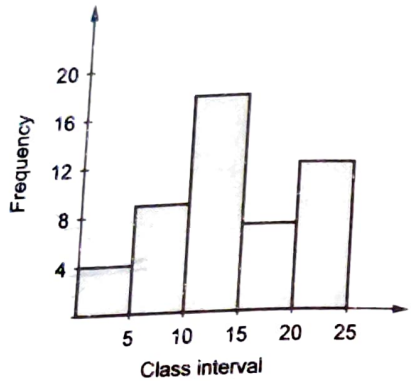


Fig. 4.6.4 Histogram

4.6.7.3 Clustering Data Reducing

- Clustering partition the objects into similar group and dissimilar group.
- Very effective if data is clustered but not if data is "smeared".
- Effectiveness of this approach is based on the nature of group data.
- If the quality of cluster is may be based on its diameter.
- If the distance between two object in cluster is centroid distance.
- Hierarchical clustering and be stored in multi-dimensional tree.

Example 1 :

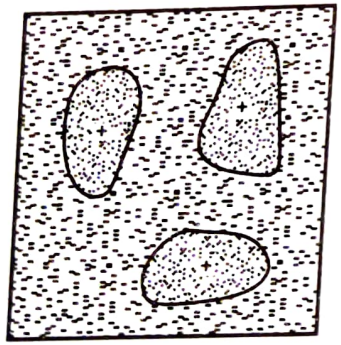


Fig. 4.6.5 Cluster data reduce

4.6.7.4 Sampling Data Reducing

- Sampling technique can be reduce large data set to be a smaller random or s of the data.

- When sampling technique in large data set to be D, contain N tuples, random sampling without replacement (SRSWOR) of size S. (S < N).

Different-Different Sampling

i) Cluster sampling
ii) Stratified sampling
iii) Simple random sampling

i) Clustering sampling : Obtained by selecting cluster from the population on the basis of simple random sampling.

ii) Stratified sampling : Obtain by several selecting random sample from every population of stratum.

iii) Simple random sampling : It is a equal probability chance of being selected.

For example 1 :

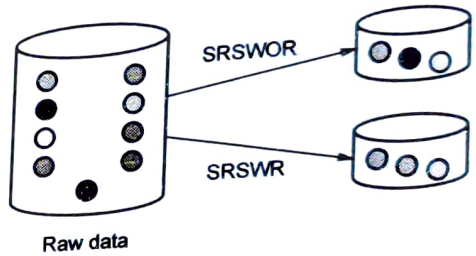


Fig. 4.6.6 Sampling data reducing

4.7 Discretization and Concept Hierarchy Generation

- It is categorize based on how to discretization is performed to efficient.
  i) Supervised ii) Unsupervised.
- Interval level can be used to change real data value.
- Discretization reduce size of future useful data.
- Changing numerous values reduce to interval level and simplifies the original data.
- There is categorical attributes have deterministic finite number of distinct values.
- This is leads to a concise, simply, knowledge-level representation of mining output.

4.7.1 Hierarchy Generation for Categorical Data

This are different-different several method are as follow :

i) Representation of a set of attribute but not partial order.

ii) Representation of a partial ordering of attribute explicitely at the schema le user.

iii) Representation specification of attribute, but not their partial order of a pro a hierarchy by explicit data mining.

For example : Concept of hierarchy expression by using prespecified semantic ne Age, Name, College name, B.E. degree, degree grade. They are linked semantically regarding the useful notation of student reco user to specify only the attribute of degree grade for a hierarchy student.

4.7.1 Type of Attribute in Discretization

i) Nominal

ii) Ordinal

iii) Continuous

i) **Nominal :** Attribute in discretization useful define as a different-different un set of value.

Example : In a class name of student.

ii) **Ordinal :** Attribute is discretization useful define as a similar order set of val

Example : Technical publication can be publish higher education book.

iii) **Continuous :** Define as a fix value.

Example : Any item cost.

4.7.2 Discretization and Concept Hierarchy Generation for Numeric Data

- Concept of hierarchy to be represent constructed simple alltomatical based on distribution analysis.
- Useful numeric concept hierarchy generation method are as follow :
  - i) Binning
  - ii) Histogram analysis
  - iii) Cluster analysis
  - iv) Entropy based discretization
  - v) Segmentation by natural partitioning

i) Binning : This methods can be smooth sorted all data value. This sorted value can be distributed into different-different number of "buckets" or bin. Because bin method can be consult the all value.

ii) Histogram analysis : Histogram is define partition rule it can be define range of value. Histogram can be applicable recursively and each and every partition in order to be create multiple concept hierechy.

iii) Cluster analysis : Clustering is define similar and dissimilar value of collection or organized in group or "cluster".

iv) Entropy-based discretization :

- Entropy based method will use to be describe class information entropy of subset partitions to select boundaries for discretization.
- When use to minimum to minimum length of description criterion.
- When description of attribute cannot be compressed more.
  - Description of splitting point ($\log_2[N-1]$ bits) +
  - Description of bins (class distribution)
- Short of few thresholds, homogeneous (Single-class) bins.
- Split worthwhile if information gain >

$$\frac{\log_2(N-1)}{N} + \frac{\log_2(3^K-2) - KE + K_1E_1 + K_2E_2}{N}$$

where E = Entropy, K = Class in original set (E, K). Subset before threshold (E_1, K_1) ter threshold (E_2, K_2).

v) Segmentation by natural partitioning :

- An employee was braking up his annual salaries in to the ranges like (50,000-100,000) are often more desirable than range like (51, 263, 89-10, 895) arrived at by using cluster analysis.
- Using segment numeric data into relatively uniform "natural" interval.
- Recursively level by level based on value range at the most significant digit.
- When recursively rule can be applied to each interval creating a concept of hierarchy for the given numeric attribute.

4.7.2.1 Discretization and Concept Hierarchy Generation for Categorical Data

i) Partial ordering of attribute explicitly at the schema level by experts :

- Easily define to concept hierarchy by specifying partial attribute at a schema level.

ii. Portion of a hierarchy by explicite data grouping :

- It is a realistic to specify explicit grouping for a small portion of intermediate data.

iii) If set of attribute but not their partial ordering :

- Its based on system can automatically generate the attribute ordering so construct to a meaningful or explination concept hierarchy.

iv) Its only a partial set of attribute :

- Handle such partially specified hierarchies.

4.8 Data Mining Primitives

A data mining task can be specified in the from of a data mining query which input to data mining system. Term of data mining query is to be define these primitives allows the user to interactively communicate with data mining system during discovery in order to direct the mining process.

i) The set of task-relevant data to be mined :

- This specifies the portion of the database or the set of data in which the user is interested include database attribute call dimension of interest.

ii) The kind of knowledge to be mined :

- This specifies the datamining function to be performed, such as characterization, discrimination, association, classification, prediction, clustering, outlier analysis etc.

iii) Background knowledge to be used in discovery of process :

- In this knowledge about the domain to be mined is useful for guiding the knowledge of discovery process and evaluating the patterns found. Where concept of hierarchy are a popular form of back ground of knowledge.

iv) Interesting measures and thresholds for pattern evaluation :

- Used to guide the mining process, after discovery, to be evaluate the discovered pattern in different-different kinds of knowledge may have different interesting measure.

v) Expected representation for visualization the discovered patterns :

- When the form in which discovered pattern are to be displayed, which may include different-different rules, graph, decision tree, table and cube.

4.8.1 Data Mining Model and Task

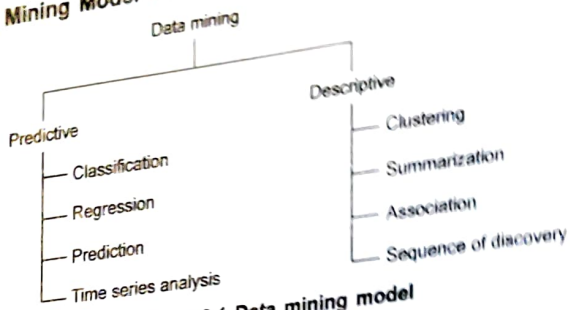


Fig. 4.8.1 Data mining model

4.8.2 Task Relevant DATA

- Data warehouse name
- Database table
- Condition for data selection
- Dimensions
- Data grouping criteria

For example : You are a manager in a Home Entertainment Company incharge of sales in the Gujarat state and M.P. state, in particular, you would like to more study or analysis the buying trends of customer in Gujarat state. Rather them mining in the entire database, now you can also specify the attributes of interest to be considered in analysis process. These are referred a most relevant attributes.

Concept type = " Home Entertainment" can represent lower level concept.

like {" TV, "CD player", "refrigerator"}.

First primitive in the specification of the data on which mining is to be performed. Normally user is using a subset of the database. It is related to immedatly mined the database, explicitly since the more then pattern generated could be ascending with respect to the or according to database size. Further more different type of the patterns found would be referred as relevant attributes. In a relation database the set of task relevant data can be collected, via query involving different type of operation projection, join and aggregation. Now this curve inhole data can be thought of as a sub task or subset of the mining task the data collection process results is define this new relation is call initial data relation, this initial data relation can be grouped or associated according to the condition specified in query this initial relation may or not corresponds to physical relation in the hole data base. Since virtual relations are called views in the field of database this set of task in related data for analysis is call mixable view

If the data task is to study association between frequently at " Home Entertainment company by customer in Gujarat, the task related data can analysis by providing following information.

- The name of database (e.g. Home Entertainment")
- The name of table (e.g. item, customer, purchases, item sold)
- The attribute are dimensions (e.g. name and price for the item table and income
- Condition for data selection (e.g. retrieve data pertaining to purchases made Gujarat for this year).
- Data grouping criteria (e.g. data retrieved be grouped by certain attribute an qu can be used to retrieved the task relevant data).

In data warehouse different type of large data are typically stored in multidimensional array structure and relation structure. But in term of task relevant da can be define by the condition based data filtering and slicing of the data cube.

4.8.3 The Kind of Knowledge to be Mined

It is important to define the kind of knowledge to be mined the data mining functi to be performed, that knowledge include concept description characterizati Discrimination. Association, classification / prediction, clustering and outerlier analy In addition to define kind of knowledge to be mined for a given data mining task, wh user can be define by using more different types of pattern templates that all discover patterns match.

4.8.4 Back Ground Knowledge to used in Discovery of Process : Concept Hierarchies

In this knowledge about the domain to be mined is useful for guiding the knowled of discovery process and evaluating the patterns found. When concept of hierarchy are popular write from background knowledge or the background knowledge allows data be mined at multiple level of abstraction.

A concept of hierarchy is explain a sequence of mapping from a set of low-lev concepts to higher - level more general concepts. Now concepts hierarchies, are a usef from background knowledge. In that they allow raw data to be handled at highe generalized levels of abstraction. Generalize of the data is achieved by replacin primitive - level data by higher level data in using hierarchy user to view the data more meaningful and explicit abstraction and makes the discovered patterns to easy t understand.

In resulting data appear over generalized, schema hierarchy also allow specializatio where concepts values are replaced by lower level concepts by using rolling up an

drilling downs user can view the data from different perspectives, gaining further into hidden data relationship mainly or mapping are typically data specific concept hierarchies can often be automatically discovered refined based the data statical distribution. For example : Your Manager Home Entertainment Company who is interested in studying habits of the customer at different location may prefer the concepts hierarchy. Now making manager, however may prefer to see location. Organization with respect to linguistic lines facilities the distribution of commercial ads.

Different type of concept hierarchies :

i) Schema hierarchy
ii) Set grouping hierarchy
iii) Operation - derived hierarchy
iv) Rule - based hierarchy.

i) Schema hierarchy :

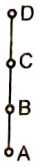
A schema hierarchy is a total or partial order among the attribute in the database schema. This hierarchy may formally express semantic relationship between attributes.

Schema hierarchy of a relation for address containing the attribute street city provinces or_state and country, now we can define a location schema hierarchy by the following total order.

Example 1 : Street < city < Province _ or _ State < Country

In this example street is at a conceptually lower level then city, which is lower than province_or_state which is conceptually lower than country.

Example 2 :



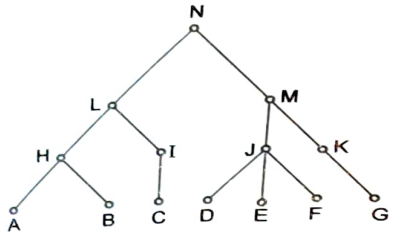
In a database more than one schema hierarchies can be created by using different sequences.

ii) Set grouping hierarchy :

Set grouping hierarchy organizes values for a given attribute or dimension into grouping of constants or range values. Total order can be defined among the groups. When set grouping hierarchy can be used to define schema hierarchies.

Example :

In this figure explain that attribute N is divided into two groups L and M. These are divided on the basis of domain of attribute N. Further partial order of the attribute L



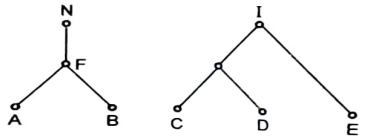
and M has been defined on set or instance or value of these attribute so, Partial ord
L and M are H, I, J and K respectively.

A set grouping hierarchy can be used for transforming a schema hierarchy or an
set grouping hierarchy to from a sophisticated hierarchy.

iii) Operation_derived hierarchy :

Operation_derived based hierarchy is defined by a set of operations on the
These operation are specified by users, experts or the datamining system. T
hierarchy are generally defined for numerical attributes. Such operation can be as sim
as range value comparison; as complex as a data clustering and data distribu
analysis algorithm.

Example :

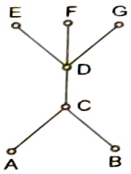


In this operation_derived hierarchy for a simple range value comparison operation
given data. The attribute H and I has two different values. The attribute with high
placed as right child and with less value is at left likewise in this figure with
subpart G has value less than I and E has greater than I and so on.

iv) Rule - based hierarchy :

A rule based hierarchy either a whole concept hierarchy or a portion of it is defin
by a set of rules and is evaluated dynamically based on the current data and r
definition. A lattic like structure is used for graphically describing this type of hierarc
in which every child - parent path is linked with a generalization rule. In this figu
highlights a rule based concept hierarchy.

Example :



For a particular data mining task some specified hierarchies may not be descriable. So, there should be mechanisms for automatic generation of concepts hierarchies based on the data distribution of concept hierarchy is totally depend on data sets.

4.9 Discretization and Concept Hierarchy

Discretization : Discretization techniques can be main use to reduce the number of values for a given continuous attribute, by dividing the attribute into a range of intervals. Interval value tabels can be used to replace actual data of values. These methods are continued most of time is spent on sorting the data at each step. The smaller the number of distinct values to sort, the faster these method should be.

Concept hierarchy : In this concept hierarchy can be used to reduce the data collecting and replacing low- level - concepts by higher level concept. Although detail is lost by such generalization its becomes meaningful and it is easier to interpret.

4.9.1 Discretization and Concept Hierarchy Generation for Numeric Data

It is more difficult to specify concept hierarchy for numeric attribute due to the wide diversity of possible data range and the frequent update if data values.

Concept hierarchy for numeric data can be constructed is automatically but is based on data analysis. These are different types of methods.

i) Binning or below - binning

ii) Histogram or histogram analysis

iii) Cluster analysis

iv) Segmentation by natural partitioning

i) Binning : These technique can be use recursively to the partition in order to generate concept hierarchies, basically in this technique attribute values can be discretized by distributing the values into bin and replacing each bin by the mean value.

ii) Histogram : Partitioning rules can be applied to define range of values.

iii) Cluster analysis : Partition data into groups in cluster analysis. Each cluster may be further decomposed into sub-cluster, forming a lower cluster in the hierarchy, cluster may also be grouped together to form a higher level.

iv) Segmentation by natural partitioning : Breaking our home loan in the range of like (Now like 10,0000 - 20, 0000) after more than range like (5000, 10,000, 150000) arrived at by cluster analysis. The 2-3-4 rule can be used to segment numeric data into relatively uniform " natural " intervals in basically the rule partitions a given range of data in 2-3 or 4 equinity intervals.

4.9.2 Discretization and Concept Hierarchy Generation for Categorical Data

Now categorical attribute have finite number of distinct values, but is distribution or categorical no ordering all the values. These are different types of methods : Specification of partial ordering of attribute explicitly at the schema level by experts :

User can easily define concept hierarchy by exploring a partial attribute at a schema level.

ii) Example : Street < City < Province < State < Country .

Specification of a portion of a hierarchy by explicit data grouping.

It is realistic to specify explicit grouping for a small portion of the intermediate level.

iii) Specification of a set of attribute but not their partial ordering :

In system try to automatically generate the attribute ordering so as to construct a useful concept hierarchy.

Review Questions

1. *Explain why data pre-processing required.*
2. *Explain data clearing.*
3. *Explain data integration.*
4. *What is the purpose of data transformation.*
5. *Explain data reduction.*
6. *Explain binning methods for data.*
   Smoothing sorted data for price in dollar 4, 8, 9, 15, 21, 2 8, 2
7. *Explain Following :*
   i) Schema hierarchy ii) Set grouping hierarchy iii) Rule

5 Concept Description and Rule Mining

Syllabus

What is concept description ? - Data generalization and summarization-based characterization - Attribute relevance - Class comparisons association rule mining : Market basket analysis - Basic concepts - Finding frequent item sets : Apriori algorithm - Generating rules - Improved apriori algorithm - Incremental ARM - Associative classification - Rule mining.

Contents

5.1 What is Concept Description ?

Association is a descriptive approach to exploring data that can help identify relationships among values in a database. Association rule mining finds interesting association rules about items that appear together in an event such as a purchase transation by ralational database and other information repositories.

For example : Association rule mining is market basket analysis, we have some large number of items like "bread, butter, eggs, cereal", customer their market basket with some subset of the items and we get to know what item people by together, even if we don't know who they are.

Market basket analysis is just one form of frequent pattern mining there are many kind of frequent pattern association rules in frequent mining can be classified in various way based on the following criteria.

i) Association values handles in the rule.

ii) Association multidimensional rule.

iii) Association on the kinds of rules.

iv) Association on abstraction rules.

v) Association on complement ness of pattern rule.

i) Association values handles in the rule : If a rules association between the presence of items base on boolean association rule.

ii) Association multidimensional rule : Rule reference two or more dimension association, for example age income and buys then it is multidimensional association rules.

iii) Association on the kinds of rules : Association mining can generate a large number of rules of which are redundant correlation relationship among itemset.

iv) Association on abstraction rules : Association rule mining can find out rules at differing levels of abstraction in the set of rule mined is call multi-level association rules.

For example : Customer rule is multi-level.

v) Association on complement ness of pattern rule : Association rule application may different requirements regarding complement ness the pattern to be mined which is turn can lead to different evaluation and optimization methods.

5.2 Data Generalization and Summarization based Characterization

Let item $I = \{I_1, I_2, I_3, \ldots, I_m\}$ be a set of M distinct attribute. Let D be a database of item, where each record of (Tuple) T has unique identifier and contains a set of items

such that $T \subseteq I$. According association A => B B where $B \subseteq I$ set of item called itemset of An, B, where A is called antecedent and B is called consequent.

Where is many important measures for association rules. According database is taken from any Bazaar (Complete shop)

Transaction_ID	ITEMSET
T_{01}	BREAD, MILK, BISCUIT, CORNFLAKES
T_{02}	BREAD, TEA, BOURNVITA
T_{03}	JAM, MAGGIA, BREAD, MILK
T_{04}	MAGGIA, TEA, BISCUIT
T_{05}	BREAD, TEA, BOURNVITA
T_{06}	MAGGIA, TEA, CORNFLAKES
T_{07}	MAGGIA, BREAD, TEA, BISCUIT
T_{08}	JAM, MAGGIA, BREAD, TEA
T_{09}	BLADE, FORMAL BRUSH
T_{10}	COFFEE, COCK, BISCUIT, CORNFLAKES
T_{11}	COFFEE, COCK, BISCUIT, CORNFLAKES, SUGER
T_{12}	COFFEE, SUGAR, BOURNVITA
T_{13}	BLADE, COFFEE, COCK
T_{14}	BLADE, SUGAR, BISCUIT
T_{15}	COFFEE, SUGER, CORNFLAKES

Table 5.2.1 Data base

i) Let's $I = \{I_1, I_2, I_3, I_4, \ldots, I_m\}$ be a set of M, distinct attributes.

ii) D be a database.

iii) D database, where each record of tuple (T).

iv) $T \subseteq I$ an association rule is an implication of the A => B.

v) $B \subseteq I$ are set of items call item set A_k and B, where A is called antecedent and B is called consequent itemset : Simply set of item I.

vi) K is a set of items.

vii) Support : Support (S) of an association rules according is the ratio of the records that contain $A \cup B$ to the total number of records in the DataBase (DB).

For example : According database the support (S) of a rules 10 % that it m that 10 % of the total records contain A ∪ B, support of a A => B is the percen of transaction in D (database) contain both A and B.

viii) Confidence : Confidence (C) is the association rule, is the ratio of the numbe records that contain A ∪ B to the number of records that contain A.

Example according database rule has a confidence (C) of 80 %, it means that 8 of the records contain A also contain B.

If the confidence of a value indicates the degree of correlation in data set A and its percentage of transaction in D containing A, that also contain in B.

xi) Threshold (σ) : Frequent item set is said to be frequent if it satisfies the minim support threshold (σ).

x) Minimum support and a minimum confidence are said to be strag rules.

5.3 Attribute Relevant (Association Mining)

Association is a data mining function that discovers essential role in many d mining task probability of the co-occurrence of items in the collection in correlation casual structure among set of item is call association rule mining to following rule.

i) Transaction

ii) Items and collection

iii) Spare data

iv) Itemset

v) Frequent-item

vi) Frequent-pattern

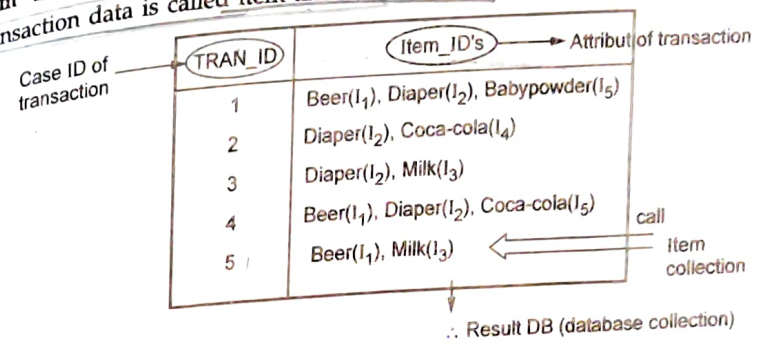
i) Transaction : Let $I = \{I_1, I_2, I_3, I_4, ..., I_m\}$ be a set of items and a transactio database $DB = < T_1, T_2, ..., T_n >$, where. T_i $(i \in [1 \cdots n])$ is a transaction whic contains a set of items in I. Every transaction has a key label it called T_{ID}.

T_{ID}	List of Item_IDs
T_{001}	I_1, I_2, I_5
T_{002}	I_2, I_4
T_{003}	I_2, I_3
T_{004}	I_1, I_2, I_4
T_{005}	I_1, I_3
T_{006}	I_2, I_3

T_{007}	I_1, I_3
T_{008}	I_1, I_2, I_3, I_5
T_{009}	I_1, I_2, I_3

Table 5.3.1 Transaction table

ii) Item and collection : A collection of item is correlation with each case in transaction data is called item and collection.



iii) Spare data : Item is not present in data-base or transaction collection its have present Null Value is called spare data.

T_{ID}	Item_IDs
1	I_1, I_2, I_5
2	I_2, I_4
3	^, ^
4	I_1, I_2, I_4
5	I_1, I_3

Null value

null value in DB

iv) Itemset : A collection of item is any combination of two or more item in a data base or transaction is call item set.

two or more data item

T_{ID}	Item_IDs
1	(Beer, Diaper), (Diaper, Babypowder)
2	(Diaper, Coca-cola) (Coca-cola, Milk, Beer)
3	(Beer, Milk, Diaper)
4	(Beer, Milk)

Called Itemset

v) **Frequent-item** : A item-set whose support count is greater than or equal to the minimum support there should special by the user in database or transaction.

T_{1D}	Item	Frequent-item
I_{001}	F, C, a, R, Q, K, m, P	F, c, a, m, P
I_{002}	a, b, c, F, Z, m, a	F, c, a, b, m
I_{003}	b, F, h, c	F, b
I_{004}	b, c, K, Q, P	c, b, P
I_{005}	a, F, C, Z, P, m, n	F, c, a, M, P

vi) **Frequent-Pattern** : To design a compact data structure for efficient frequent-pattern mining.

5.3.1 Class Comparisions Association Rule Mining

Several algorithms have been proposed in the literature to address to problem of mining assocation rules which seem to be the most popular in many application for enumering frequent itemsets to be following algorithms.

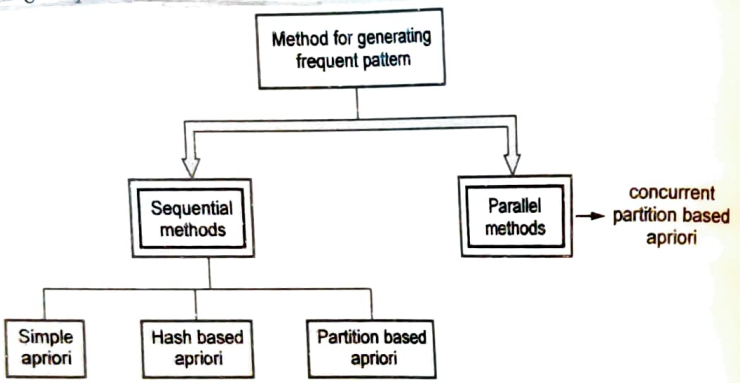


Fig. 5.3.1 Class comparisons association rule mining

1) **Sequential methods** : Sequential algorithm also forms the foundation of most known algorithm whether sequential, if uses a monotone property stating.

i) K-itemset to be frequent K-1 itemset have to be frequent. The use of its fundamental property.

ii) A sequence database consist of ordered elements.

reduce the computational cost of candidate frequent itemset generation in cases of extremely large input set with outsized frequent 1-itemset.

K-Sequence database

T_{1D}	Itemset
T_{001}	I_1, I_2, I_4
T_{002}	I_1, I_3, I_4
T_{003}	I_1, I_4, I_5
T_{004}	I_2, I_5, I_6

→

S_{1D}	Sequence
T_{001}	$<I_1 (I_1 I_2 I_3) (I_1 I_3) I_4 (I_3 I_6)>$
T_{002}	$<(I_1 I_4) I_3 (I_2 I_3) (I_1 I_6)>$
T_{003}	$<(I_5 I_6) (I_1 I_2) (I_4 I_6) I_3 I_2>$
T_{004}	$<I_5 (I_1 I_6) I_3 I_2>$

Sequential method still suffers from the two main problems.

i) Repeated I/O scanning

ii) High computational cost

2) **Simple apriori** : Simpal Apriori algorithm major hurdle observed with most real dataset is a two process. Following two process can Apriori algorithm according to develop by Agrwal R and Srikant R.

i) **The join step** : To find L_K a set of candidate K-item set is generated by joining L_{K-1} item itself. Rules of Join : order the item so you can compare item by item the join of L_{K-1} is possible only if its first $(K-2)$ item are common.

ii) **The prune step** : The join step will produce all K-item sets, but not all of them are frequent. So scan database to see when join step produce and empty set to be following process of Apriori algorithm.

3) **Hash based apriori** : A hash based technique can be applied so that it reduce the size of the candidate K-item set in C_K for $K > 1$. We have improve our main aim is to reduce the number of scans on the database.

For example : When we scanning each transaction in the dataset to generate the frequent 1-item L_1 from the candidate 1-itemset C, we can generate all of 2-itemset for each transaction hash them into corresponding bucket counts in the hash table.

H (A, B) = ((Order of A) * 10 + order of B) mod 8.

Bucket address	0	1	2	3	4	5	6	7
Bucket count	3	2	2	4	2	2	4	4
Bucket contents	{A, D}	{A, D}	{A, E}	{B, C}	{B, E}	{B, D}	{A, C}	{A, B}
	{A, D}	{A, D}	{A, E}	{B, C}	{B, E}	{B, D}	{A, C}	{A, B}
				{B, C}			{A, C}	{A, B}
				{B, C}			{A, C}	{A, B}
	{C, E}			{B, C}				

4) **Partition based apriori** : A partition of the database refer to any subset of transaction contained in the database D. Partition reduces the number of datab scans to it divides the database into small patri such that each partition can handled in the main memory. Partition scan DB only twice.

Scan 1 : Partition database and find local frequent pattern.

Scan 2 : Consolidate global frequent pattern initially the database D is logic partitioned into n partitions.

Phase I : Read the entire database once takes n iteration.

Input : P_i where i = 1 — n

output : Local large itemset of all length

Merge phase :

Input : Local large item sets of same lengths from all n partition.

Output : Combine and generate global candidate itemsets.

Note : The set of global candidate item sets of length J is Jumputed.

Phase II : Read the entire database again and all so takes n iteration.

Input : P_{iB} where i = 1.....n;

Output : Counter for each global candidate itemset and count their support.

Item sets that have the minimum global support along with their support an again reads the entire database.

Example :

T_{ID}	Items
10	I_1, I_2, I_3
20	I_2, I_3, I_4
30	I_2

Item	Support	Unit
I_1	50 %	Yes
I_2	100 %	Yes
I_3	50 %	Yes
I_4	25 %	No

5) **Parallel methods** : Parallel algorithm implemented over distributed memor system. Parallel work can be followed.

i) Each processor gathers the locally frequent itemset of all size in one pas over their local database. Then all potentially frequent itemset are then broadcast to other processor.

ii) Other then each processor gathers the counts of these global candidates itemsets. There are two major approaches for using processor.

a) Distributed memory machines

b) Shared memory processor system

5.4 Market Basket Analysis (Basic Concepts)

Market basket analysis is a modelling technique based upon customer can be collection of item purchased by a customer in a single transaction, it call market baset.

Or

One basket tell only interested in the different type of items purchased.

Basic concept of market basket :

For $I = \{I_1, I_2, I_3, ... I_n\}$ a set of item, transaction T is a set of items $T \subseteq I$ and T is a transaction database $T = (t_1, t_2, ..., t_m)$

Market basket transaction

t_1 = {bread, cheese, milk}

t_2 = {apple, eggs, salt, butter}

t_3 = {bread, butter, beer}

t_4 = {pizza}

t_5 = {soda, salad}

⋮ ⋮

t_m = {biscuit, milk, eggs}

Concepts

I : set of all item sold in the store

Transaction : item purchased in a baset it call T_{1D}

The model rules of basket

i) A transaction t contain A a set of item in I if $A \subseteq t$

ii) An association rule is an implication of the $A, B \subseteq I$ an $A_m, B = \phi$

iii) An item set is a set of item A = {bread, cheese, milk} is a-itemset

iv) K-itemset is an itemset A = {bread, butter, beer}

5.4.1 Use of Market Basket

1) Retail : Customer purchases different set of products, different quantities, different time, different place.

ii) Mobile calling : Different-different customer can calling different num different company calling.

iii) Reservation : Customer/Different person can be get reservation different-different place in different time.

iv) Credit card : Use customer to ATM machine of different bank, differ different place.

v) Medical insurance claim : Complications based on certain combinat treatment.

5.4.2 Market Analysis

i) Association rules can be applied on other types of "baskets"

ii) Banking product used by custome according analysis (loan, deposit, investment, purchase).

iii) A store entity provides information about product.

5.5 Finding Frequent Itemsets

Frequent items will play a role in the frequent-pattern mining it is necessa perform one scan of transaction database DB to identify the set of frequent ite compact data structure can be designed based on the following observation,.

i) Pattern a set of items/subsequences, substructures that occurs frequently data set.

ii) The set of frequent item of each transaction can be stored in some cor structure.

iii) Multiple transaction share a set of frequent items, it may be possible to m the shared sets with the number of occurrences registered as count.

iv) If two transaction share a common prefix, according to some sorted orde frequent items.

v) Support and confidence.

a) Support count : The support count of an item A is denoted by A. Count dataset T is the number of transaction in T that contain X. Assume T ha transactions.

$$\text{Support} = \frac{(A \cup B) \cdot \text{count}}{\eta}$$

b) Confidence : The rule hold in T with confidence % of transaction that cont A also contain B.

$$\text{Conf} = P_i (B/A)$$

$$\text{Confidence} = \frac{(A \cup B) \cdot \text{count}}{A \cdot \text{count}}$$

5.5.1 Motivation of Frequent Patterns

i) It may be possible to avoid repeatedly scanning the original transaction database.

ii) If the frequent items are sorted in their frequency descending order.

iii) Frequent pattern can be automatically classify web document.

iv) Frequent pattern finding inherent regularities in database sysetm.

v) A large database is compressed into a condensed, smaller data structure.

5.5.2 Application of Frequent Pattern

i) Market basket analysis

ii) Web document analysis

iii) Multiple marketing

iv) Catalog design

v) Multimedia frequent pattern

5.5.3 Frequent Pattern Mining Important

i) Association rule mining

ii) Frequent pattern based clustering

iii) Relational database

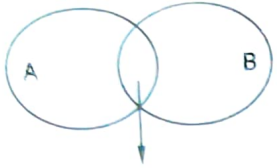
iv) Data warehouse

v) Patter mining

5.5.4 Use Association Rule in Frequent Pattern

Transaction-id	Item set
1	P q R
2	P R S
3	P S T
4	q T U
5	q R S T U

Here will be sup_{min} = 50 % and $conf_{min}$ = 50 %.
Frequent pattern
{P = 3, q = 3, s = 4, T = 3, ps = 3}



$S = \frac{(A \cup B)\text{.count}}{n}$

$C = \frac{(A \cup B)\text{.count}}{A\text{.count}}$

Assocation rule mining

$P \rightarrow S$ (60 % and 100 %)

$S \rightarrow P$ (60 % and 75 %)

Fig. 5.5.1 Use assocation rule in frequent pattern

5.6 Apriori Algorithm

Apriori algorithm assumes that items within a transaction or itemset are sorted in lexicographic order (based on item name). This order provides a logical manner in which itemset can be generated and counted. It employs an iterative approach known as a level-wise search. Where (K – 1)- itemsets are used to explore K-itemset.

Apriori algorithm

A two-step process–

1) **The join step** : Find L_K the set of candidate of K-item set, join L_{K-1} with itself

   // **Rule for Joining** : Order the items first so you can compair item by item the of L_{K-1} is possible only if its first (K – 2) items are in common.

2) **The prune step** : The "join" step will produce all K-item sets; but not all of t are frequent. Scan DB (database) to see when join step produces and empty set

Pseudo-code :

Apriori algorithm

Function count (C : a set of itemsets, D : database)

1) begin
2) For each transaction $T \in D = UD_i$ do begin
3) For *al* subsets $A \subseteq T$ do
4) if $A \in C$ then
5) A · count ++;
6) end
7) end

Method of Apriori

1) Input : Set of all large (K – 1)-itemset L_{K-1}
2) Output : A superset of the set of all large K-itemset

Join step

1) I_i = item i
2) Insert into C_K
3) Select $P \cdot I_1, P \cdot I_2, P \cdot I_3 \ldots P \cdot I_{K-1}, q \cdot I_{K-1}$
4) L_{K-1} is P, L_{K-1} is q
5) $P \cdot I_1 = q \cdot I_1$ and .... $P \cdot I_{K-2} = q \cdot I_{K-2}$ and $P \cdot I_{K-1} < q \cdot I_{K-1}$

Pruning step

1) All itemset $C \in C_K$ do
2) All (K – 1) subset S of C do
3) If ($S \notin L_{K-1}$) then
4) delete C from C_K

Algorithm : Apriori

//Apriori algorithm proposed by Agrwal R, Srikant R.

Procedure

1) $C_1 = I$; //candidate t-itemset
2) Generate L, by traversing database and counting each occurrence of an attribute in a transaction
3) For ($K = 2$; $L_{K-1} \neq \phi$; K++) do begin
   //Candidate itemset generation
   New K-candidate itemset are generated from (K – 1) large itemset.
4) C_K = apriori-gen (L_{K-1})
   //Counting support of C_K
5) Count (C_K, D)
6) $L_K = \{C \in C_K | C\text{.count} \geq \text{minsup}\}$
7) end
8) L : = U_K L_K

Example : Apriori algorithm

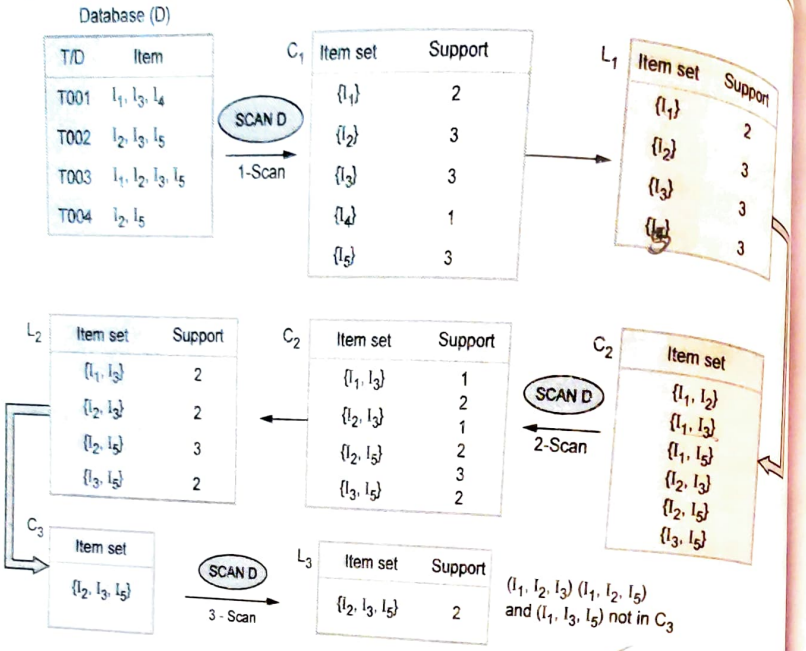


Fig. 5.6.1 Apriori 1 example

Another example of apriori algorithm

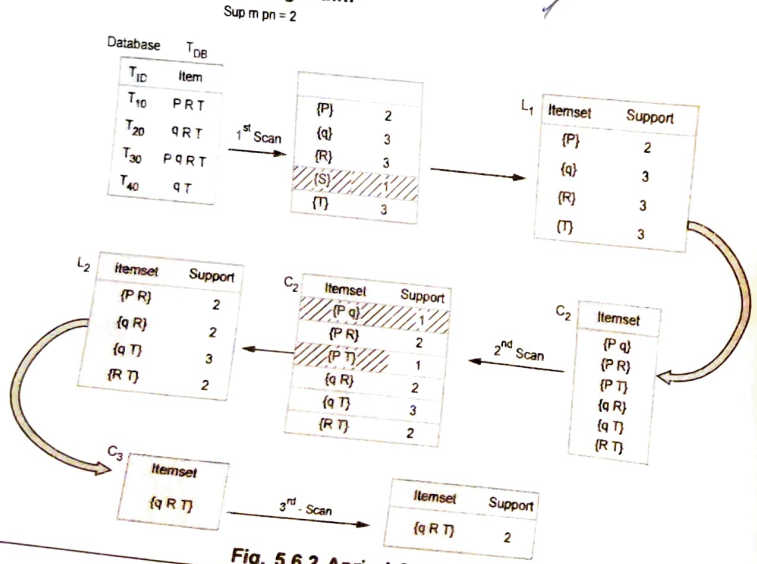


Fig. 5.6.2 Apriori 2 example

5.6.1 Apriori Algorithm Generating Rules

1) Only considering large itemset of the previous pass. The number of candidate large itemset is significantly reduced in the first pass.
2) Itemset with only one item are counted, it can discovered large itemsets of the first pass are used to generate the candidate sets of the second pass using.
3) Candidate itemsets are found, their support are counted to discover the large itemsets of size two by scanning the database in third pass.
4) Large itemset of the second pass are considered as the candidate set to discover large itemset of this pass.
5) Iterative process terminates when do new large itemset are found.

Step 1 : 1-item can generating frequent pattern.

i) Each item is a member of set of candidate.
ii) The set of frequent 1-itemset, L_1, consist of the candidate 1-itemsets satisfying to minimum support (depend only minimum sup).
iii) Then next

Step 2 : 2-item can generating frequent pattern.

i) The set of frequent 2-itemset, L_2 the algorithm use to L_1 join L_1 to generate a candidate set of 2-itemset, it call C_2.
ii) Next
iii) The transaction in D (1-scan) are scanned and the support count for each candidate item set in C_2 is accumulated.
iv) The set of frequent 2-itemset, L_2 is then determined, consisting of those candidate 2-itemset in C_2 having minimum support.

Step 3 : 3-itemset can generating frequent pattern

i) The generation of the set of candidate 3-itemsets, C_3, according use of the Apriori algorithm property.
ii) Find C_3 to order, we have compute L_2 join L_2.
iii) $C_3 = L_2$ Join L_2
iv) Join step is complete and Prune step will be used to reduce the size of C_3.
v) Prune step helps to avoid heavy computation due to large C_K.

Step 4 : 4-itemset can generating frequent pattern

i) Accounting algorithm use L_3 join L_3 to generate a candidate set of 4-itemset, C_4. The join result (I_1, I_2, I_3, I_5).

ii) This item set is Pruned since its subset is (I_2, I_3, I_5) is not frequent.

iii) Where use to strong assocation rule satisfy both minimum support δ' mimimum confidence.

Step 5 : Association rules from frequent pattern can be generated.

i) For each frequent itemset "1" generate all non-empty subset of *I*.

ii) For every non-empty subset S of 1, output rule "S => (1 – S)"

Support_count(1)/support_count (S) > min_conf

(Where, min_conf = minimum confindence threshold)

iii) Confidence is defined as the measure of certainty assocation with each discovered pattern.

If A => B this is called minimum confidence threshold.

$$\text{Confidence } (A \Rightarrow B) = \frac{\#\_\text{tupies}\_\text{containing}\_\text{both}\_A\_\text{and}\_B}{\#\_\text{tuples}\_\text{containing}\_A}$$

5.6.2 Drawback of Apriori

1) Multiple database scans are costly.

2) Mining long pattern needs many passes of scanning and generates lots of candidates.

5.6.3 Frequent Itemset Generation of Apriori

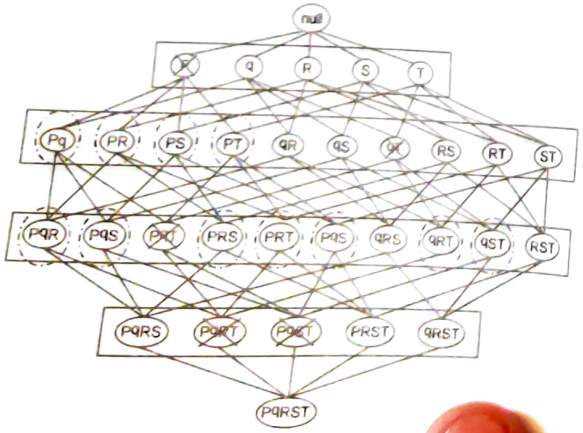


Fig. 5.6.3 Frequent itemset generati... thm

5.6.4 Architecture of Apriori Algorithm

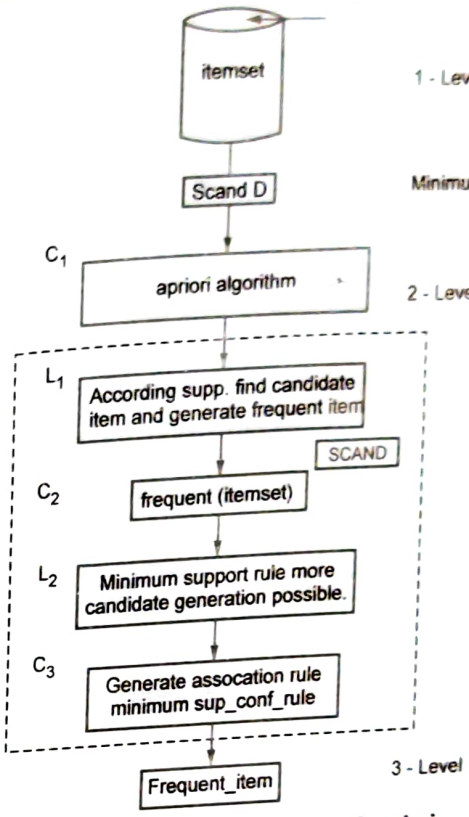


Fig. 5.6.4 Architecture of apriori

5.7 Improved Apriori Algorithm

Mining frequent pattern in transaction databse, time-series database and many other kinds of database has been studied popularly in data mining. The previous studies adopt an Apriori-like candidate set generation and test approach. Candidate set generation is still costly, especially when there exist large number of patterns. This drawback of apriori can be improved. Following improvement (Apriori needs n + 1 scans)

1) Hash-based item
2) Transaction based
3) Partition based
4) Dynamic itemset counting
5) Pincer-search algorithm
6) Sampling

1) Hash-based item : The main drawback of frequent-pattern itemset is that they are very large in number to compute and store computer to present leads to the introduction of closed frequent itemset and maximal frequent itemset in has based technique can be applied so that it reduce the size of the candidate. K-itemset C_k for $K > 1$.

For example : When we scanning each transaction in the dataset to generate the frequent 1-itemset L_1 from the candidate 1-itemset C_1 we can generate all of 2-itemset for each transaction has item into corresponding bucket count in the hash table.

H (A, B) = ((Order of A) * 10 + Order of B)) mod 6

B-Address	1	2	3	4	5	6
B-count	3	2	3	2	4	2
Basket	{A C}		{A C}	{A B}	{A C}	{A D}
Count	{A E}	{A D}	A B	{A B}	{A C}	{A B}
	{E E}	{A E}	{A B}		{A B}	
			{A B}		{A E}	

2) Transaction based : Purpose of improve apriori to reduced transaction does not contain frequent K-item can not be next frequent of (K + 1). So improve apriori algorithm.

3) Partition based : Partition based apriori can reduce the number of database scan to, it divide the database into small partition such that each partition can be handled in the main memory. Let the partition P of the database $R_1, R_2 .... R_P$ can be following scan in database.

Scan I : It to finds the local large itemset in each partition R_i. $\{A \mid A \cdot count \in S\} \cdot |R_i|$ the local large itemset can be store minimum time local memory L_i, can be found by using a level-wise algorithm such as apriori.

Scan II : It uses the property that a large itemset in the whole database must be locally large in at least one partition of the database.

Then the union of the local large itemset found in each partition are used as the candidate and counted through the whole database to find all the large itemset.

For example :

T_{ID}	Itemset
T_1	Bread, Butter, Beer
T_2	Butter, Beer, Milk
T_3	Butter
T_4	Bread, Butter

Scan D

L^1 = {{Bread}, {Butter}, {Beer}, {Bread, Butter}, {Bread, Beer}, {Butter, Beer}, {Bread, Butter, Beer}, {Milk}, {Butter, Milk}, {Beer, Milk}, {Butter, Beer, Milk}}

L^2 = {{Butter} {Bread} {Bread, Butter}}

C = {{Bread}, {Butter}, {Beer} {Bread, Butter}, {Bread, Beer}, {Butter, Beer}, {Bread, Butter, Beer}, {Milk}, {Butter, Milk}, {Beer, Milk}, {Butter, Beer, Milk}} $\xrightarrow{\text{scan D}}$ count support

L = {{Bread} {Butter}, {Beer}, {Bread, Butter}, {Butter, Beer}}.

4) Dyanamic Itemset Counting (DIC) : DIC to generate and count the itemsets; it reduce the number of database scan, database is viewed as intervals of transaction are scan sequentially. While scanning the first interval 1-itemset are generated and counted at the end of first interval, 2-item which are potentially large are generated. While scanning the second interval all 1-itemset and 2-itemset generate are counted at the end of second interval 3-itemset that are potentially large are generated are counted scanning the third interval together.

In general K^{th} interval the (K+1)-itemsets which are potentially large are generated and counted it together with the previous itemset in the later interval. If can be rewinds the database to the beginning and counts the itemset which are not fully counted. DIC itemset are marked in four different way as they are counted.

- Solid box (☐) : An itemset we have finished counting and exceeds the support threshold minsupport.
- Solid circle (◯) : We have finished counting and it is below minimumsupport.
- Dashed box : ⬚ An itemset we are still counting that exceeds minimumsupport.

Solid : exceed minimum support
Dash = below

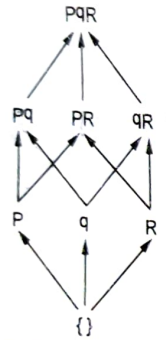
- Dashed circle ◌: An itemset we are still counting that is below minimumsupp

For example :

T_{ID}	Itemset		
	P	q	R
10	1	1	0
20	1	0	0
30	0	1	1
40	0	0	0

According transaction database lattice

1)



1) Fig. 5.7.1 (a)

2) Before any transaction : Itemset lattice

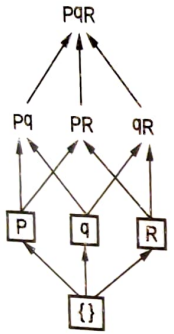


Fig. 5.7.1 (b)

Counter will be

P = 0, q = 0, R = 0

According DIC Empty itemset in solid box.

3) After this transaction lattice.

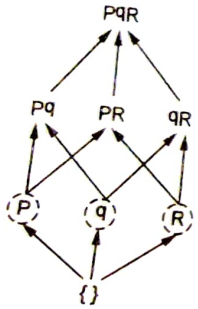


Fig. 5.7.1 (c)

In this DIC process : itemset we finish counting of 1-itemset then store memory.

4) After counting P = 2 and q = 1 and R = 0, where Pq = 0 here we have change Pq to dash box their counter are be greater than minimum-support is at present 1 and add a counter for Pq because both of its subset are boxes.

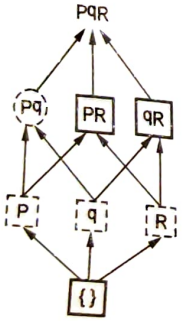


Fig. 5.7.1 (d)

5) After previous transaction can be read

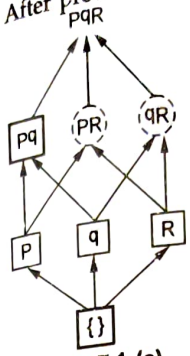


Fig. 5.7.1 (e)

i) After this process counter : P = 2, q = 2, R = 1. Here will be according DIC Pq = 1, PR = 0 and qc = 0.

ii)Here will be Pq has been counted is to satisfies/satisfied minimumsupport we can change it to solid box. Then we change here at qR to a dashed box.

6) After previous transaction lattice according

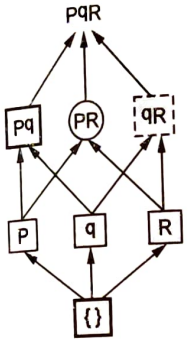


Fig. 5.7.1 (f)

i) After this process can be P = 2, q = 2, R = 1, Pq = 1, PR = 0 and qc = 0.

ii) Here PR and qR has been counted and PqR are donot counted because their one subset of circle, according DIC.

Disadvantage of DIC :

i) Very long process can be.

ii) Very costly process.

5) Pincer-search : Pincer-search is two-way approach make use of both the properties and speed up the search for maximum frequent set.

Pincer-search algorithm begins with generating 1-itemset as apriori algorithm but uses top-down search to prune candidate product in each pass. This is done with the help MFCS set. MFS denote set of maximal frequent set, storing all maximally frequent itemsets found during the execution.

Anytime during the execution MFCS is a superset of MFS

i) $\text{FREQUENT} \subseteq \{2^A \mid A \in \text{MFCS}$

ii) $\text{INFREQUENT} \subseteq \{2^A \mid A \in \text{MFCS}$

For example :

T_{ID}	Itemset
10	Butter, Coke, Beer
20	Beer, Chips
30	Coke, Butter
40	Beer, GroundNuts
50	Coke, GroundNuts

Step I : $L_0 = \phi$, K = 1;

C_1 = {{Butter}, {Coke}, {Beer} {Chips} {GroundNuts}

MFCS = {Butter, Coke, Beer, Chip, GroundNuts}

MFS = ϕ

Pass I : DB (database) is read to count support

Item	Support
Butter	2
Coke	3
Beer	3
Chips	1
GroundNuts	2

{Butter, Coke, Beer, Chip, GroundNuts} → 0 after

MFCS = {Butter, Coke, Beer, Chips, GroundNuts}

MFS = ϕ :

L_1 = {{Butter}, {Coke} {Beer}, {Chip} {GroundNut}}

$S_1 = \phi$

Here will be $S_1 = \phi$ at preset don't need to update MFCS.

C_2 = {{Butter, Coke}, {Butter, Beer}, {Butter, Chips}, {Butter, GroundNut}
{Coke, Beer}, {Coke, Chip} {Coke, GroundNut}, {Beer, Chip},
{Beer, GroundNut} {Chip, GroundNuts}

Pass II : DB (database) is read and count support of element in C_2 and MFCS.

Item	Support
Butter, Coke	2
Butter, Beer	1
Butter, Chip	0
Butter, GroundNut	0
Coke, Beer	1
Coke, Chip	0
Coke, GroundNut	1
Beer, Chip	1
Beer, GroundNut	1
Chip, GroundNut	0
Butter, Coke, Beer, Chip, GroundNut	0

Here will be MFS = ϕ

L_2 = {{Butter, Coke}, {Butter, Beer}, {Coke, Beer}, {Coke, GroundNut},
{Beer, GroundNut}, {Beer, Chips}

S_2 = {{Butter, Chip}, {Butter, GroundNut} {Coke, Chip}, {Chip, GroundNut}

after this

{Butter, Chip} in S_2 for {Butter, Coke, Beer, Chip, GroundNut} in MFCS,

MFCS = {Butter, Coke, GroundNut} {Coke, Beer, Chip, GroundNuts}

{Butter, GroundNut} in S_2 {Coke, Beer, Chip, GroundNut} in MFCS

Then {Butter, GroundNut} is donot include in MFCS then {Butter, Coke, Beer, Groundnut} in MFCS we got

New element in MFCS {Butter, Coke, Beer} and {Coke, Beer, GroundNuts} Then {Coke, Beer, GroundNut} is all ready in MFCS it is excluded from MFCS.

MFCS = {{Butter, Coke, Beer} {Coke, Beer, Chip, GroundNuts}}

after {Coke, Chip} in S_2. Then

MFCS = {{Butter, Coke, Beer} {Coke, Beer, GroundNut} {Chip, Beer, GroundNut}

Then {chip, GroundNut} in S_2. Then

MFCS = {{Butter, Coke, Beer}, {Coke, Beer, GroundNut} {Chip, Beer}}.

after we generate candidate set

C_3 = {{Butter, Coke, Beer} {Chip}}

after this according algorithm to stop, because other no more candidate set.

6) Sampling : If basic idea of sampling to be efficient counting of itemset this algorithm can be reduce scan of large database so Apriori algorithm can be use to sampling algorithm. These are following two phase of scan.

1) Potentially Large (PL) itemset and used as candidate to be counted using the entire database.
2) Other candidate set (according minimum support) by applying union Border set of function $(BD)^U$.
3) Scan data store C.
4) $C = BD^U$ (PL) UPL is a scan function.

Example :

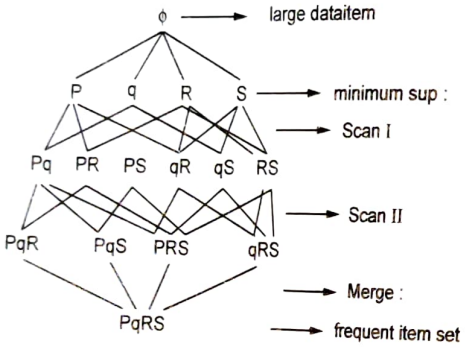


Fig. 5.7.2

5.8 FP Growth Algorithm

Let $I = \{I_1, I_2, ...I_m\}$ be a set of item and transaction database $T = \{T_1, T_2, T_3, ...T_n\}$ then $T_i = (I \in [I...n])$ is a transaction which contains a set of items I. Support is occurrence frequency A in database, A is a frequent pattern if the support A is no less than a prede-fined minimum support call threshold (-), the problem of finding the complete set of frequent pattern is call frequent pattern mining problem.

F-P growth tree can be design based on the following observation.

1) It to be perform one scan of database to be identify the set of frequent item.
2) Store in item after frequent, this process can be avoid repeatedly scanning database.
3) Multiple transaction share an identical frequent itemset, after they can be merged into one with the number of occurrences registered as count.

4) The same node is encountered in another transaction, increment the support count of the common node or call prefix. After this process can be generated tree traversal, because point to its occurences in the tree or chain of node-link.

5.8.1 Construction of FP-Tree : FP-Growth Method

1) Create or consist of one root labeled as always "Null" if set of item prefix sub trees as the children of the root node, then frequent-item is it Header.
2) Each node in the prefix sub tree consist of three fields
   a) item-name b) Count c) node-link
3) Then entry in the frequent-item header table consist of two fields
   a) item-name b) Head of node-link
4) Always each item transaction are processed in L order.

5.8.2 FP - Growth Method : Example

T_{ID}	Item
T_{10}	A, B, E
T_{20}	B, D
T_{30}	B, C
T_{40}	A, B, D
T_{50}	A, C
T_{60}	B, C
T_{70}	A, C
T_{80}	A, B, C, E
T_{90}	A, B, C

i) Count minimum support according you but here will be minimum support 2(are required). After minimum support count database.

Item	Support
A	6
B	7
C	6
D	2
E	2

According support 2/9 = 22 %

ii) I-scan of database is same as apriori often than support count.

iii) Frequent item is sorted order of descending support.

Item	Support
B	7
A	6
C	6
D	2
E	2

iv) Then resulting L = {B = 7, A = 6, C = 6, D = 2, E = 2}

v) Then generate tree root node is "null".

vi) Each node prefix subtree consist fields item-name, supp.count, node link.

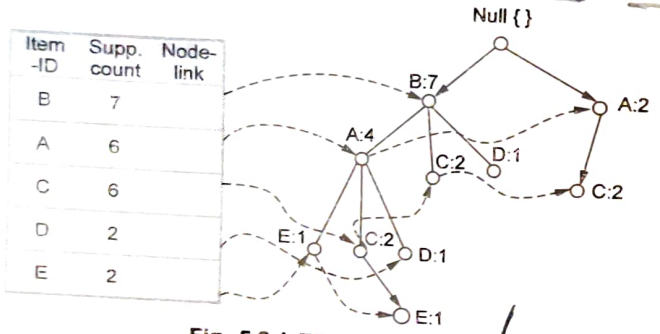


Fig. 5.8.1 FP growth tree

5.8.3 FP-Tree by Creating Conditional Pattern Tree

i) Start from each frequent length 1-pattern

ii) Conditional pattern based of prifix path of FP-tree is based on suffix pattern.

iii) FP-tree suffix pattern generated conditional pattern.

Item	Conditional pattern	Conditional FP-tree	Frequent pattern
E	{(B, A : 1), (BAC : 1)}	(B : 2, A : 2)	BE : 2, AE : 2, BAE : 2
D	{(BA : 1), (B : 1)}	(B : 2)	BD : 2
C	{(BA : 1), (B : 2), (A : 2)}	(B : 4, A : 2, A : 2)	BC : 4, AC : 2, BAC : 2
B	{(B : 4)}	(B : 4)	BA : 4

E as suffix base : E-conditional pattern 2 corresponding prefix path.

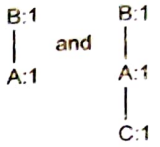


Fig. 5.8.2

5.8.4 Benefit of FP-Growth Tree

i) FP-tree avoid costly or repeated database scans.

ii) FP-tree based mining adopts a pattern-fragment growth method to avoid the costly generation of a large number of candidate sets.

iii) FP-growth eliminate repeated database scan.

iv) FP-growth is an order of faster than apriori.

v) Easy to handle large database as compared apriori.

5.9 Increment Association Rule Mining

Association rule mining problem to find out all the rules in the form of A => B, where A and BCI are set of item, called itemset, association rule mining is use two step process.

i) **The join step :** Find db K the set of candidate of K-item set, join db-1 with itself.

//Rule for joining :

Order the items first so you can compare item by item the join of db-1 is possible only if its first db-2 item are in common.

ii) **The prune step :** All the association rules that have value exceed a minimum confidence threshold. There process of K-itemset but not all of them are frequent.

When new transaction are innserted into the database, then new association rules and some existing association rule would become invalid prior. Algorithm may be reapplied the mining the whole increment data when then database has been changed.

This process is very costly even if small amount of new transaction is into a database to generate a new rule or incremental rule of mining can maintenance of frequent itemsets involves searching for two kinds of itemsets.

i) Lower set : Frequent itemset that became infrequent.

ii) Higher set : Infrequent itemset after increment data to the database.

One incremental approach fast update (FUP) is based on the Apriori algorithm, each iteration K, scan both db and D from the (db is update database) with candidate generated from the prior iteration, can be scan large item in D, the difference is that the number of candidate examined at each iteration is reduced through pruning of the candidate.

For example :

Item	Type
A	Wheat
B	Rice
C	Juice
D	Beer
E	Milk
F	Fruit

Transaction

	T_{ID}	Itemset					
DB	1	A	B				
	2	A	B	C			
	3		B	C		E	
	4			C	D		F
	5		B			E	
	6	A					
	7		B				
	8	A	B	C			F
	9	A				E	
	10	A	B	C	D	E	

Table 5.9.1 Database

5.10 Associative Classification Rule Mining

Classification is one of the most important task in data mining. Classification method that integrates association rule mining into classification problem. Associative classification achieves high classification accuracy, its rules are interpretable and it provides confidence probability when classifying objects which can be used to solve classification problem of uncertainty. Associative classification is classifying using these phases.

i) Rule generation phase : Use association rule mining technique to find database transaction item in frequent patterns containing classification.

ii) Building classifier : Using the generated and redundant class assocation rules.

iii) Classification : Using some experiments done over Associative Classification (AC) with the help of classification.

T_{ID}	Itemset
T_{100}	P,q,R
T_{200}	P,R
T_{300}	P,S
T_{400}	q,T,U

Item	Support
P	3
q	2
R	2
P,R	2

Min support = 2
min confidence = 2

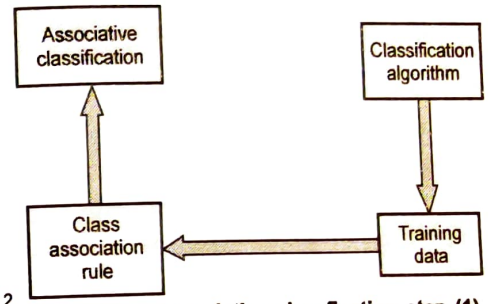


Fig. 5.10.1 Associative classfication step (1)

OR

Rule of P => R

i) Support = Support {A ∪ C} = 2 or 50 %

ii) Confidence = Support {A ∪ C} | support A

$$= \frac{3+2}{2} = 2.5$$

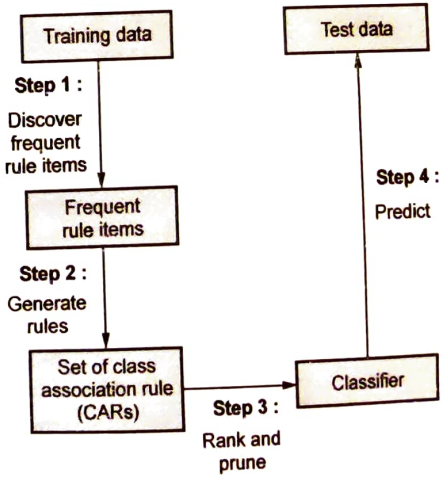


Fig. 5.10.2 Associative classfication step (2)

5.10.1 Advantages of Associative Classification

i) Redundant rule and compact set guarantee.

ii) Generate small set of rules.

iii) Depend only support and confidence.

5.10.2 Disadvantages of Associative Classification

i) Associative classification can provide more rule because first read assocation rul than classified.

ii) Redundant rules may also be included in the classifier which increase the time when classify data.

5.10.3 Association Rule of Confidence Support

Let A be an itemset $A = \{T_1, T_2, T_3 .... T_n\}$. A transaction T is said to contain A if and if only $A \subseteq T$.

Association rule is implication of the form A => B where ACI and BCI, are itemset, $A \cap B = \phi$.

Association rule $A \Rightarrow B$ hold with support (S), where S is percentage of transaction in D that contain $A \cup B$. This contain is call $P(A \cup B)$.

Association rule A => B is confidence (C) in the transaction D, where (C) is percentage of transaction D containing A that also contain B, this is probability of P(B/A).

$$S(A => B) = P(A \cup B)$$

$$C(A => B = P(B/A) = \frac{S(A \cup B)}{S(A)}$$

when

$$\text{Support} = \frac{(A \cup B) \cdot \text{count}}{n}$$

$$\text{Confidence} = \frac{(A \cup B) \cdot \text{count}}{A \cdot \text{count}}$$

For example :

Transaction ID	Item
100	m, n, o
200	m, o
300	m, p
400	
500	n, q, R

Item	Support
M	3
N	1
O	2
P	1
q	1
R	1

5.10.4 Associative Classification

Associative Classification Problem

AC is a special case of association rule discovery in which only the class attribute is considered in the rule's right-hand side (consequent); for example, in a rule such as $X \rightarrow Y$, Y must be a class attribute. One of the main advantages of using a classification based on association rules over classic classification approaches is that the output of an AC algorithm is represented in simple if-then rules, which makes it easy for the end-user to understand and interpret it. Moreover, unlike decision tree algorithms, one can update to tune a rule in AC without affecting the complete rules set, whereas the same task requires reshaping the whole tree in the decision tree approach. Let us define the AC problem, where a training data set T has m distinct attributes A1A2,...,Am and C is a list of classes. The number of rows in T is denoted $|T|$. Attributes can be categorical continuous. In the case of categorical attributes, all possible values are mapped to a set of positive integers. For continuous attributes, a discretization method is used.

Solution scheme and example

An AC task is different from association rule discovery. The most obvious difference betwen Association rule discovery and AC is that the latter considers only the class attribute in the rules Consequent. However, the former allows multiple attribute values in the rules consequent.

Table 5.10.1 shows the main important difference between AC and association rule discovery, where over fit-ting prevention is essential in AC, but not in association rule discovery as AC involves using a subset of the discovered set of rules for predicting the classes of new data objects. Over fitting often occurs when the discovered rules perform well on the training data set and badly on the test data set.

Differences between Association rule discovery and Associative Classification

The main differences between AC and association rule discovery

Sr. No.	Association rule discovery	Associative classification
1.	Unsupervised learning	Supervised learning
2.	Is to discover associations between items in a transactional database. There could be more than one attribute in the consequent of a rule.	Is to construct a classifier that can forecast the classes of test data objects. There is only attribute (class attribute) in the consequent of a rule.
3.	Over fitting not an issue.	Over fitting is an important issue.

Table 5.10.1

6 Classification and Prediction

Syllabus

What is classification and prediction ? - Issues regarding classification and prediction : Classification methods : Decision tree, Bayesian classification, Rule based, CART, Neural network. Prediction methods : Linear and nonlinear regression, Logistic regression. Introduction of tools such as DB Miner /WEKA/DTREG DM tools.

Contents

6.1 What is Classification and Prediction

i) Classification :

Classification is the process of assigning an object to a catagory according to its characteristics of the object.

Classification is to be classify before object analysis a set of pre-classified data to learn a model that can be used to classify unseen data object into one of imported or useful predefined class.

ii) Prediction :

Prediction is the process according to the class which the unseen pre-defined class of object which have already been analyzed throughout our classification process, then we can simply object class the statical techniques of regression to make prediction of continuous values is call prediction.

6.2 Issue Regarding Classification and Prediction

6.2.1 Classification

Predicts categorical class labels is based on the training set and the value in a classifying attribute and uses it classifying new data.

1) Test sample is compared with the classified result based on training set and new data.
2) Accuracy rate is the percentage of test sample that are based on classification.
3) Test set is independent of training set, otherwise data will be over-fittting concurrently occurance.
4) Use of classify data tuple whose class labels are not known.

Process

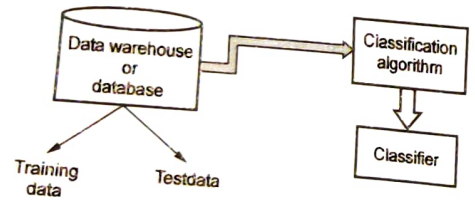


Fig. 6.2.1 Classification process

6.2.1 Evaluating Classification

1) Accuracy : Classification accuracy is two type process.
   i) Classifier accuracy : Based on predictor class label.
   ii) Predictor accuracy : Based on only guessing values of related attribute according.
2) Speed : Based on only for traning time to.
3) Robustness : Only for missing values.
4) Scalability : Efficiency in store database or data warehouse.

6.2.2 Prediction

The predicated class label is a categorical attribute to classification is clearly useful in many decision problem.

1) Predicts unknown values.
2) Data cleaning : Preprocess data in order to reduce noise and handle missing value.
3) Relavance analysis : To remove the irrelevant redundant attribute.
4) Data transformation : Data can be normalize.

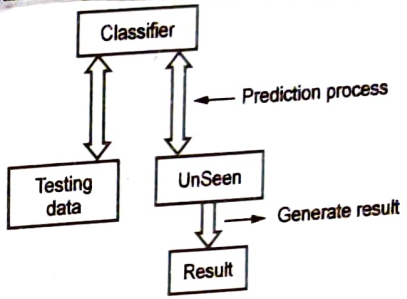


Fig. 6.2.2 Prediction

6.2.2.1 Application of Prediction

1) Credit approval
2) Target marketing
3) Medical diagnosis
4) Fraud dection
5) Data mining prediction area

6.3 Various Classifiers and Classification Methods

6.3.1 Decision Tree

A decision tree is tree structured prediction model where each internal node denoted a test on an attribute, decision tree is outgoing branchess can be represt outcome of test and each leaf node is labeled with a class distribution.

Construction decision tree

i) Starting form a root node representing the whole training data.

ii) Data is split into two or more subset based on the values of an attribute, is called hear child node or child.

iii) Each subset a child node is created subset is associated with the child node.

iv) Separately repeated on the data in each of child node.

v) Terminate criterion is satisfied.

Training set data :

B_T	B_{T2}	B_{T3}	B_{T4}	Pass ϕ (P)
B_1	B_2	B_3	B_4	Y
B_1	B_2	B_3	C_4	Y
B_1	C_2	B_3	B_4	Y
B_1	C_2	C_3	C_4	N
B_1	D_2	B_3	B_4	Y
B_1	B_2	B_3	C_4	N
C_1	C_2	C_3	C_4	N
D_1	C_2	C_3	C_4	N

Decision Tree According Training Data

According Decision Tree ($B_{T1} = B_1, B_{T2} = C_2, B_{T3} = A_3$) is sort to the use node : B_{T1}, B_{T2}, B_{T3} will be decision classify instance of represent tree.

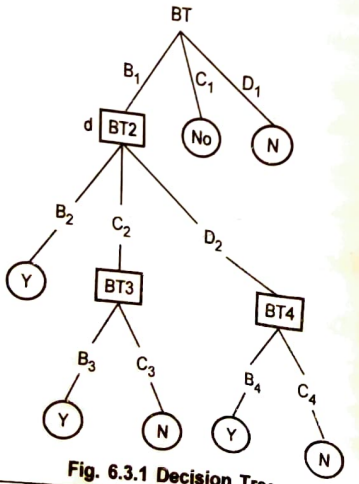


Fig. 6.3.1 Decision Tree

6.3.1.1 Algorithm for Decision Tree

- **Basic algorithm :**
  - i) Tree is constructed in a top-down recursive divide-and-conquer according.
  - ii) Add all training data in root.
  - iii) Categorical attribute according data.
  - iv) Partitioned recursively based on selected attributes.
  - v) Test all attribute are selected on the basis of heuristic search data.
- **Condition for stop partition attribute :**
  - i) All samples for a given node belong to the same class.
  - ii) There are no remaining attibute for further partition.
  - iii) There are no sample in left.

6.3.1.2 Information Gain

i) First select attribute with its highest information.

ii) Probability is P $\{I_1, I_2, ..., I_n\}$ is item set in database D is tuple T and class C.

iii) Expected information

$$INFO\ (T) = \sum_{j=1}^{B} \frac{|T_j|}{|T|} \times I\ (T_1)$$

iv) After partition tuple (T) to gain

$$Gain\ A = INFO\ (T) - INFO\ (T)$$

v) Information Gain (Probability P_i) and Tuple (T). Being class (C_i) : Total gain

$$INFO\ (T) = -\sum_{i=1}^{n} P_i \log (P_i)$$

6.3.1.3 Advantage of Decision Tree

i) GUI based : Decision tree use a graphic approach to compare competing our problem solution.

ii) Efficiency : Complex problem to easy solve and changing input information affect a decision alternatives.

iii) Revealing : Problem can be solve numerical form so easy outcome expected value of problem.

iv) Divide Task : A decision tree can divide our task to easy understand all attribute.

v) Structure : After root node same as binar tree.

vi) Pruning : This technique remove some branch of the tree after constructed is to prevent tree over fitting the training data.

Example :

Calculate information gain of following problem

Sr. No.	p	q	Class
1	1	0	Y
2	1	1	Y
3	1	1	Y
4	1	0	N
5	1	1	Y
6	0	0	N
7	0	0	N
8	0	0	N
9	1	1	N
10	1	0	N

Ans. : Mainly according problem : solution

i) Class lable of problem two (i) Y (ii) N is call n = 2.

ii) When class C_1 is called Y and class C_2 is called "N" then $C_1 = 4$ and $C_2 = 6$.

iii) Hence root node (N) created tuple T.

$$\text{Information Gain (T)} = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

$$\text{N for (T)} = \frac{-4}{10}\log_2\left(\frac{4}{10}\right) - \frac{6}{10}\log_2\left(\frac{6}{10}\right) = -0.4\times(-1.3219) - 0.6\times(-0.7369)$$

$$= 0.52876 + 0.44214 = 0.970 \text{ bits}$$

$$\frac{-P}{P+N}\left(\log_2\left(\frac{P}{P+N}\right)\right) - \frac{N}{P+N}\left(\log_2\left(\frac{N}{P+N}\right)\right)$$

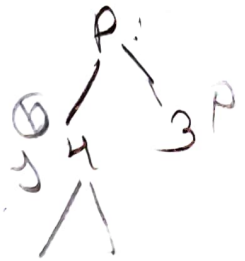
Information requirement Gain P

Gain (p)

Tuple p	1	0
P	4	3

and Gain (p_n)

Tuple q	1	0
q	4	6



$$\text{Gain p} = \text{Infor (T)} - \text{Infor p (T)}$$

$$\text{Infor}_A(T) = \frac{7}{10}\times\left(-\frac{4}{7}\log_2\left(\frac{4}{7}\right) - \frac{3}{7}\log_2\left(\frac{3}{7}\right)\right) + \frac{3}{10}*\left(\frac{-0}{3}\log_2\left(\frac{0}{3}\right) - \frac{3}{3}\log\left(\frac{3}{3}\right)\right)$$

$$= 0.7\,(0.4612 + 0.5237) = 0.689 \text{ bits}$$

Then $\text{Gain p} = \text{Infor (T)} - \text{Infor}_p(T) = 0.970 - 0.689 = 0.281$ bits

Information on Gain q.

$$\text{Gain q} = \text{Infor (T)} - \text{Infor}_q(T)$$

Tuple q	1	0
q	4	6

$$\log_2 25$$

$$\log_2 x = \frac{\log x}{\log 2}$$

$$\text{Infor}_q(T) = \frac{4}{10}\times\left(-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) + \frac{6}{10}\times\left(-\frac{1}{6}\log_2\left(\frac{1}{6}\right) - 5/6\log_2\left(\frac{5}{6}\right)\right)$$

$$= 0.4\times(0.3112 + 0.5) + 0.6\,(0.4396 + 0.2191)) = 0.3244 + 0.3898 = 0.714 \text{ bits}$$

$$\text{Gain q} = \text{Infor (T)} - \text{Infor}_q(D) = 0.970 - 0.714 = 0.256 \text{ bits.}$$

6.3.2 Bayesian Classifier

Bayesian classifier are statistical classifier it can be consist of the estimation. Pre-defined probability according distribution of each attribute belong to a particular class.

Bayesian classifier mutually exclusive and exhaustive independent attribute of class is naive classifier, based on Bayesian theorem.

Bayesian Classifier Classification are as follows

i) Bayesian classifier is defined by a set C of classes and a set x of attributes.

ii) A generic class belong to C is denoted by C_j and generic attribute A is denoted by x_i.

iii) Database is D, set of n item.

iv) $n\left(\frac{x_{ik}}{C_j}\right)$ where $x_i = \{q_1, q_2 \cdots q_i\}$ of k item in class C_j.

v) $P(x_{ik}/C_j) = \frac{n(x_{ik}/C_j)}{\sum n(x_{ik}/C_j)}$

vi) $P(C_j) = n(C_j)/n$

vii) Only estimate based frequency :

i) Prior $P(x_{ik}/C_j)$

a) Add x_j with probability class C_j.

b) x_{jk} is the number of probability x_i value with x_{ik} belong class C_j.

viii) Then $P(x_{ik}/C_j) = \frac{(x_{jk} + n(x_{jk} + C_j))}{(x_j + n\,C_j)}$

ix) Predict/Prior $P(C_j) = \frac{(x_j + n(C_j))}{x + n}$

x) Complete classify belong to new cases according, predict or prior belong $P(C_j/C_k)$ calculate.

i) $P(C_j/x_1k) = P(x_{1k}/C_j)P(C_j)/S_p(x_{1k}/C_h)P(C_h)$

ii) $P(C_j/x_{1k}, x_{2k} \ldots x_{ik}) = P(x_{2k}/C_j)P(C_j/x_{1k})/S_P(q_{2k}/C_h)$
$P(C_h/x_{1k}) \ldots./(x_{ik}/C_h)P(Ch/x_{n-1k})$

6.3.2.1 Advantages of Bayes/Bayesian

i) That is only simplified easy to process of model.

ii) Solve complete database.

iii) Problem database divide in small class continue.

6.3.2.2 Dis-advantages of Bayes/Bayesian

i) Bayesian technique is donot solve continuous data to easy handel.

ii) Bayesion technique is very difficult

iii) Problem divide small class.

iv) Result is always probability dependent.

v) Best for large database.

6.3.3 Rule Based Classifiers

A rule-based classifier is a technique for classifying records us... lection of if...than rule.

Jrip Rules : Example

Scheme : weka.classifiers.rules.JRip-F 3-N 2.0-0 2 - S 1

Relation : hepatitisdata

Instance : 39

Attribute : 20

AGE
SEX
STERIOD
ANTIVIRALS
FATIGUE
MALAISE
ANOREXIA
LIVER_BIG
LIVER_FIRM
SPLEEN_PALPABLE
SPIDERS
ASCITES
VARICES
BILIRUBIN
ALK_PHOSHATE
SGOT
ALBUMIN
PROTIME
HISTOLOGY
Class

Test mode : 10-fold cross-validation

=== Classifier model (full training set) ===

JRIP rules :

======

1. (ALBUMIN <=3.82) and (MALAISE = yes) => Class=DIE (14.0/2.0)
2. (BILIRUBIN >= 2) => Class=DIE (2.0/0.0)

Correctly Classified Instances 30 76.9231 %

Incorrectly Classified Instances 9 23.0769 %

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.826	0.313	0.792	0.826	0.809	0.764	LIVE
0.688	0.174	0.733	0.688	0.71	0.764	DIE
Weighted Avg.	0.769	0.256	0.768	0.769	0.768	0.764

=== Confusion Matrix ===

a b < – classified as

19 4 | a = LIVE

5 11 | b = DIE

Decision tree :

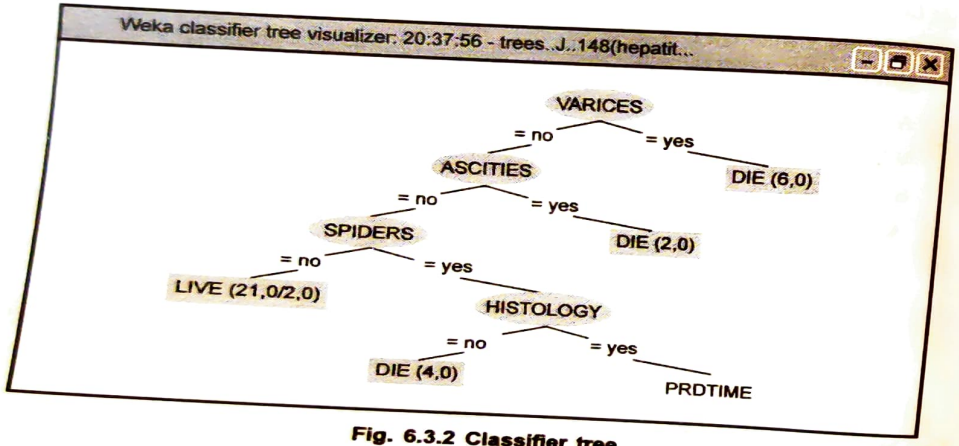


Fig. 6.3.2 Classifier tree

=== Run information ===

Scheme : weka.classifiers.trees_ J48 - C 0.25-M 2

Relation : hepatitisdata

Instances : 39

Attributes : 20

AGE
SEX
STEROID
ANTIVIRALS
FATIGUE
MALAISE
ANOREXIA
LIVER_BIG
LIVER_FIRM
SPLEEN_PALPABLE
SPIDERS
ASCITES
VARICES
BILIRUBIN
ALK_PHOSHATE
SGOT
ALBUMIN
PROTIME
HISTOLOGY
Class

Test mode : 10-fold cross-validation

=== Classifier model (full training set) ===

S48 pruned tree

VARICES = no
| ASCITES = no
| |SPIDERS = yes
| | |HISTOLOGY = no : DIE (4.0)

```
| | | | HISTOLOGY = yes
| | | | | PROTIME <= 38 : DIE (2.0)
| | | | | PROTIME > 38 : LIVE (4.0)
| | | ASCITES = yes : DIE (2.0)
VARICES = yes : DIE (6.0)

Number of Leaves : 6

Size of the tree : 11

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          32      82.0513 %
Incorrectly Classified Inastances        7      17.9487 %
=== Detailed Accuracy By Class ===
```

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.913	0.313	0.808	0.913	0.857	0.834	LIVE
0.688	0.087	0.846	0.688	0.759	0.834	DIE
Weighted Avg.	0.821	0.22	0.823	0.821	0.817	0.834

```
=== Confusio Matrix ===
a b <- - classified as
21 2 | a = LIVE
5 11 | b = DIE
```

6.3.4 CART (Classification and Regression Tree)

CART (classification and Regration tree) divide up the database Holl data into two sub-set (Right and left) so that the record within each subset are more homogeneous than in the previous one, one side of node can be analyze separately. It call recursive process again of those two subject is split again the process is repeated until the homogeneous process as same to analyze saperately again and another stopping criterion

is to the spliting could continue until all cases are perfectly classified. CART is following basis generate tree.

i) **Split selection method** : Performances towards split criteria depend on only size weighted sum of losses. It can be split = crit(split LR) = $w_L + w_R$ loss R.

ii) **Cross validation** : Every tree start to root node it can be split tree and applying to the prediction of observation from randomly.

iii) **Weight** : Weight for aggregated data set in classification problem is related to minimizing cost.

iv) **Misclassification** : Misclassification class by majority of vote and estimate the misclassification rate by proportion of the other class of database.

v) **One-sided-purity** : Single pure bucket and class we replace weighted average.

$$Crite_{LR} = \min\left(P_L^A, P_L^B, P_R^A, P_R^B\right)$$

When equivalent

$$Crite_{LR} = \min\left(P_L^A\ P_L^B, P_R^A\ P_R^B\right)$$

When split left and right bucket

$P_L^A + P_L^B = 1$ and $P_R^A + P_R^B = 1$ where probability of A and B.

vi) **Gain index :**

$$gini\ (T) = j - \sum_{i=1}^{I} P(i|T),$$

$$GG\ (T, X, Q) = gini\ (T) - \sum_{j=1}^{n} P(q_j p(x)|T| \cdot gini\ T_j$$

viii) **Decision tree :** Decision tree domain according arrange attribute.

Spliting criteria will left and right side according.

i) Criterion R_1 : **One-sided purity**

$$Crit\ L_R = \min\left(\sigma_L^2, \sigma_R^2\right)$$

Heare will be σ^2 = small, μ = Extreme

ii) Criterion R_2 : One-sided extremes

$$Crite\ L_R = \max\left(\mu_L, \mu_R\right)$$

Hear will be σ^2 = small, μ = Extrance

iii) Find low or minimum

Crite $L_R = \min(\mu_C, \mu_R)$

For Example :

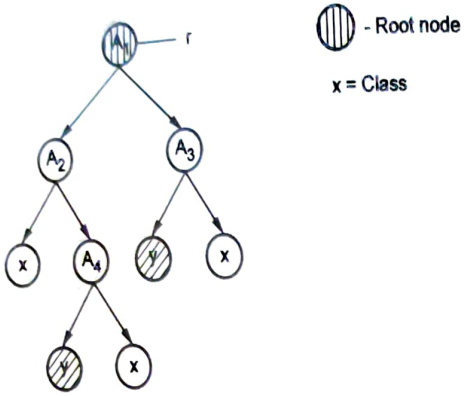


Fig. 6.3.3 Example

6.3.4.1 Advantage of CART

i) Classification as simplicity of result.

ii) Tree can be split to easy decision.

iii) CART solve any type structures.

6.3.4.2 Disadvantage

i) CART select optimal tree.

ii) Tree structure is unstable.

iii) CART based on class · tree structure.

6.3.5 What is Neural Network

Algorithm based on neural network have a lot of application in knowledge of engineering. A neural network always start with input layer in a interconnected circuit of a assembly very large number of predictor variable node it can be parallel directed links that node are neurally based, every and each pridictor variable node operate or associate only self information. These each input layer node are directly

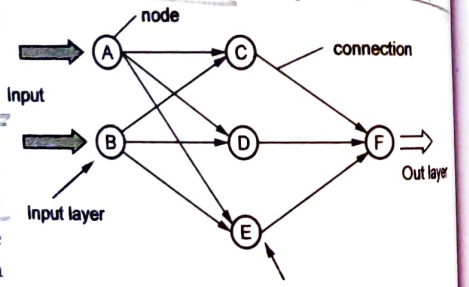


Fig. 6.3.4 Neural network

connected to a each of hidden layer node and hidden layer each node. All so connected each of outlayer node. Following neural network architecture :

The architecture of neural network here number of node and hidden layer the software must use to the number of hidden node in hidden layer. Neural network adopt various learning mechanism these are following.

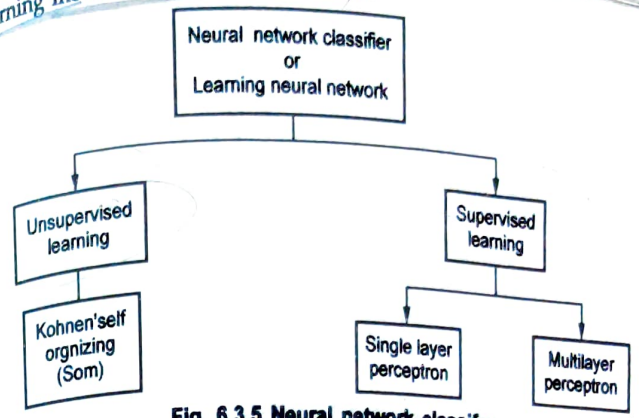


Fig. 6.3.5 Neural network classifer

i) **Learning neural network :**

i) Learning is the process neural network use to assimilatias of interconnection weight among neurons.

ii) For the neural network to be learn to improve our knowledge.

ii) **Supervised learning :**

i) Supervised training require the system developer tell the network what is the correct answer is calculated and weight adjusted such a way that once given the input.

ii) Supervised learning include error correction learning.

iii) Main aim of supervised learning to determine a set of weight which can be minimises or minimum error or error free.

iv) Which same time known common to many learning squar paradigns is call (Lms) least mean.

iii) **Unsupervised Learning :**

i) Only for external based upon only local information.

ii) This network is many time refer to use self organization network of a particular problem.

iii) Unsupervised learning is to categorize according to input patterns into a finite class.

iv) Unsupervised learning is off-line learning.

6.3.5.1 Type of Learning Rules

i) Habbian learning rules

ii) Delta learning rules

iii) Kohonen learning rules

iv) Hopfield learning rules.

i) Habbian learning rules : Different-Different interpretations of Hebbian rules

a) Weight_{mn} (New) – Weight_{mn}(old) Δ $f_{mn} \times f_{mn}$

b) Weight_{mn} (New) – Weight_{mn}(old) Δ $f_{mn} \times A_{mn}$

c) Weight_{mn} (New) – Weight_{mn}(old) Δ $f_{mn} \times B_{mn}$

Here will be w = weight, f = output, A = input, B = input.

ii) Delta learning rules : Delta learning is finite derivation of the transfer function add to the continuous equation.

finite derivation of transfer fun is add

when

Weight_{mn} (New) – Weight_{mn}(old) = f_m error.

iii) Kohonen learning rules :

i) Kohonen is unsupervised learning process.

ii) Kohonen learning rules depend only training.

iii) Its depend only self organization learning rules.

Weight_{mn} (New) – Weight_{mn}(old) = f

iv) Hopfield learning rules :

i) It is specified to weakening.

ii) Based on input and output activities.

iii) More than powerful in Habb's learning rule.

6.3.5.2 Layer of Neural Network

i) ADALINE

ii) MADLINE

iii) Perceptron.

i) ADALINE :

i) The perceptron learning rules use the output of the threshold function is (– 1 or + 1) for learning.

ii) Base on supervised learning or LMS (least mean square).

iii) Useful to describe only perceptron training rule is the way the output of the system.

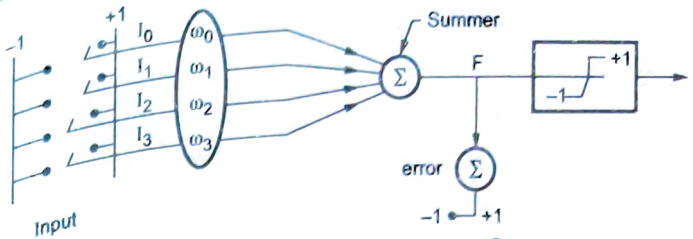


Fig. 6.3.6 Adaline

ii) MADLINE :

i) It is multiple or many AdLine.

ii) It is combination connection of AdLine.

iii) Count weight after transfer to bipolar threshold unit.

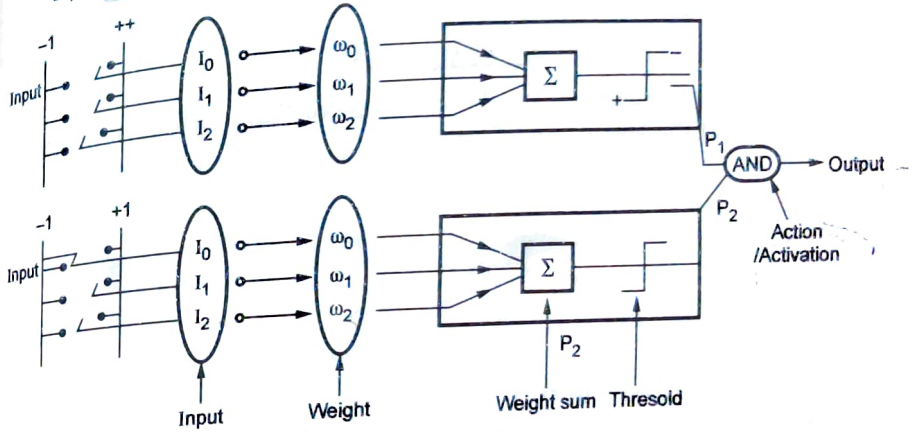


Fig. 6.3.7 Madline

iii) Perceptron :

i) Perceptron is basically represent mathematical model of biological neuron.

ii) Perceptron basically calculate the weight sum of input values.

iii) Perceptron is basically electrical signal are represented as numerical values.

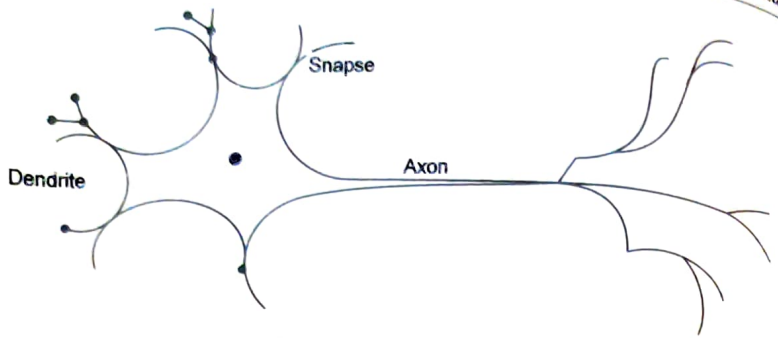


Fig. 6.3.8 Perceptron

Neuron Network use to biological neuron base.

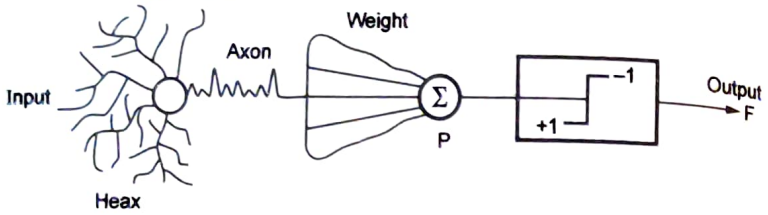


Fig. 6.3.9 (a) Neuron

Heare A and B is basically input value certain perception is call P, then weight of A and is x and y. Total weight or sum of XA + YB.

$$P = XA + YB$$

where is Z is threshold then output

$$P = \begin{cases} 1 & xA + yB > z \\ 0 & Ax + yB < z \end{cases}$$

6.3.5.3 The Back-Propagation Algorithm

Backpropagation is a neural network learning algorithm we must adjust the weight of each unit in such a way that the error between the desired output and the actual output can reduced the weight associate with it. Back propagation algorithm is a followed by Least Mean Squared (LMS) algorithm that modifies or modification are made in the backwards direction form the output layer through each hidden layer down to the hidden layer that is called backpropagation that is uses to supervised learning in which the network is trained using data for which input as well as desired output the algorithm is given as follows.

Backpropagation algorithm steps :

i) Initialize weight (to small random ≠ s) and biases in the network.
ii) Propagate the input forward.
iii) Output of an input is its actual input value.
iv) Backpropagate the error by updating weight and biases.
v) Compute the error with respect to the next higher layer.
vi) Terminating condition when error is minimum than.

6.3.5.4 Multilayer Feed Forword Neural Networks

i) The input to the network corresponded to the attribute measured for each training tuple.
ii) Multilayer feed forward consist of an input layer, output layer and also hidden layer.
iii) The number of hidden layer is arbitrary.
iv) The neural network feed-forward in that none of the weight cycles back to an input unit of a previous layer to be full connected then next forward.
v) They are then weighted and feel simultaneously to a hidden layer.

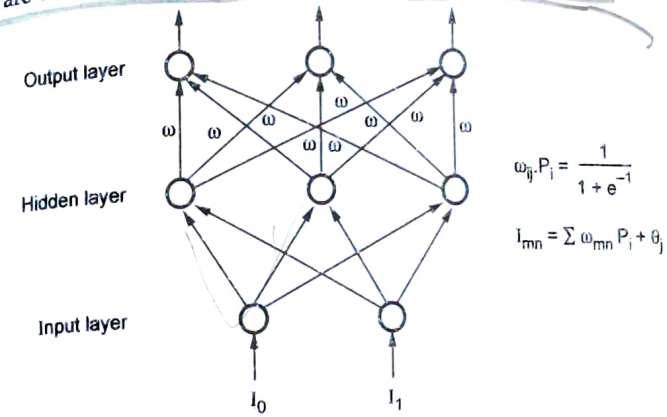


Fig. 6.3.10 Multilayer feed forward NN

Multilayer Perceptron

i) It is called MLP
ii) Multilayer perceptron can be use more than one hidden layer.
iii) Input layer represent the new information.
iv) Input layer and output layer do not directly connected in between hidden layer.

more complex things.

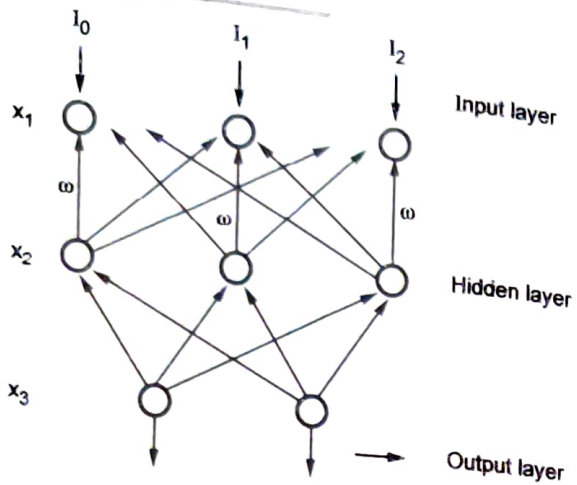


Fig. 6.3.11 Multilayer preceptron

6.3.5.5 Artificial Neural Networks

i) Neural networks is information composed large number of simple process element it called neurons.

ii) Neural network use AI to solve biological neural system.

iii) Artificial neural network process is very fast process.

iv) For example AI using human brain learns.

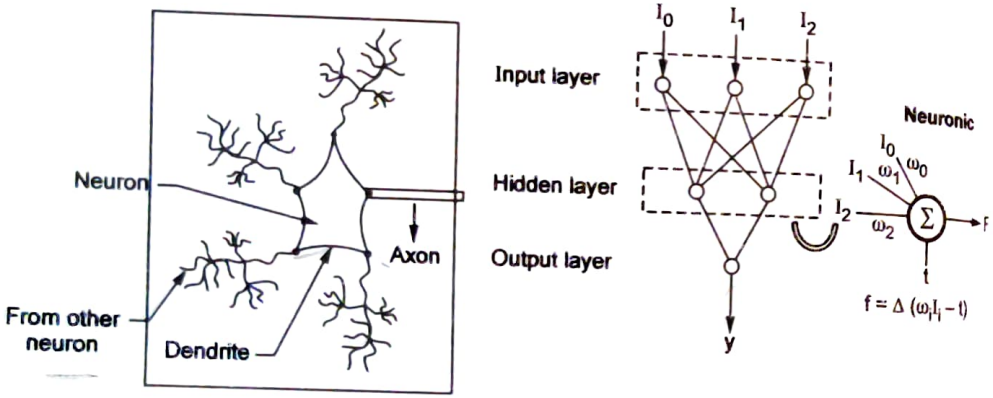


Fig. 6.3.12 ANN

Application of Neural Network

i) Neural Network in Financial

- Bank failure
- Debit risk
- Marketing
- Credit evaluation
- Future prediction
- Economic and financial forcosting
- Business success and failure prediction
- Target marketing
- Sales forcasting.

ii) Neural Network in medicine

i) Include information on symtoms

ii) Patient related diagnosis and treatment.

iii) Human resource

i) Predicting employees performance and behaviuor

ii) Determine personal resource requirement.

Artificial Neural Network Application

Artificial neural network are undergoing to the knowledge is represented and change that occurs when the present environment are as following application of neural network.

i) Language processing

ii) Pattern recognizations

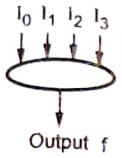
iii) Traveling salesman

iv) Character recgonization

v) Applet or image data compression.

i) **Language processing :** These application are basically speech conversion to text to speech, audio to machine and audio to video these process can automatically use to natural language processing

ii) Pattern recognitions : Neural network a different-different pattern recognization application have been use but basically neuron have two mode of pattern one is training mode and using mode, training mode neuron can be trained for a particular input pattern after this neuron using mode when a taught input pattern is detected or fach at the input neuron.



Training + Use = Result

Fig. 6.3.13 Pattern recognitions

For example : Any digital image from a electronic camera these include several and different angles of front and back, this image can be compair after using basically biggest different real or training image to using image.

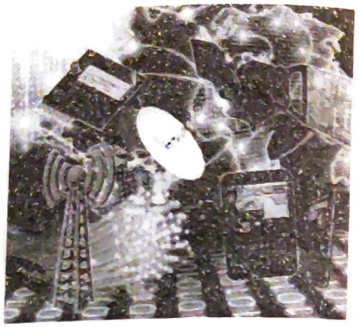


Fig. 6.3.14 Pattern recognitions

iii) Travelling salesman problem : Basically travelling salesman problem : Given a number of cities on plane and find the shortest path through which one can visit all of the cities.

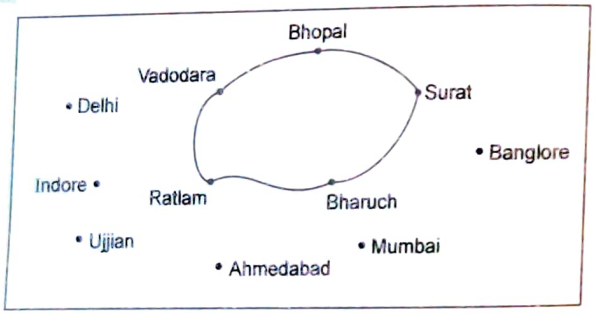


Fig. 6.3.15 Travelling cities

According to choose a random city each time and pull the point on the net that is closset to the city towards the city.

iv) Character recognization : Character regonization is another area in which neural network are providing solutions. These are very simply process can recognize hand printed characters through scanner, like a credit card, bank account debit fund, application form and put system after to recognized character in to a store database.

v) Applet or image data compression : Neural network can perform real-time compression and decompression technique this technique suitable for solving the image compressing problem.

Fig. 6.3.16 Data compression

Advantage of Neural Network

i) The neural network improve learning performance.
ii) Neural network trained rather than technique.
iii) Neural network using increase any problem scabability and efficiency.
iv) Neural network performance can applicable to better performance in large and dynamic training data.
v) Neural network can be implemented in parallel hardware.
vi) Neural network can be update with fresh data, making useful for dynamic environments.
vii) Neural network are very flexible with respect to incomplete, missing and noisy data.
viii) When an element of the neural network fails it can continue without any problem by their parallel nature.

Disadvantages of Neural Network

i) Neural network performance always depend weight value.
ii) Neural network are difficult to unstand user.
iii) Input value allways must be numerical methmatical.
iv) Testing
v) Verification

Real Life use of Neural Network

i) Frud dection

ii) Insurance

iii) Bankruptcy prediction

iv) Telecommunication

v) Loan approvals

vi) New product analysis

vii) Signature and bank note varification.

Using Data Set Neural Network Example

Viewer

Relation: hepatitis data

No.	AGE Numeric	SEX Nominal	STEROID Nominal	ANTIVIRALS Nominal	FATIGUE Nominal	MALAISE Nominal	ANOREXIA Nominal	LIVER_BIG Nominal	LIVER_FIRM Nominal	SPLEEN_PALPABLE Nominal	SPIDERS Nominal	ASTITES Nominal	VARICES Nominal	BILIRUBIN Numeric	ALK_PHOSPHATE Numeric	SGOT Numeric	ALBUMIN Numeric	PROTIME Numeric
1	30.0	male	no	no	no	no	no	no	no	no	no	no	no	1.0	85.0	18.0	4.0	?
2	50.0	female	no	no	yes	no	no	no	no	no	no	no	no	0.9	135.0	42.0	3.5	?
3	78.0	female	yes	no	yes	no	no	yes	no	no	no	no	no	0.7	96.0	32.0	4.0	?
4	31.0	female	?	no	yes	no	no	yes	yes	no	yes	yes	no	1.43	105.33	85.68	4.0	?
5	23.0	female	yes	no	no	no	no	yes	no	no	no	no	no	1.0	105.33	85.68	3.85	?
6	22.0	male	yes	yes	yes	no	no	yes	no	no	no	no	no	0.9	48.0	20.3	3.85	?
7	39.0	female	no	yes	yes	yes	no	yes	yes	no	no	no	no	2.3	280.0	98.0	4.2	?
8	62.0	female	no	no	yes	yes	no	yes	yes	no	no	no	no	1.0	105.33	60.0	3.8	?
9	41.0	male	yes	yes	yes	yes	yes	yes	no	no	no	no	no	0.7	61.0	53.0	3.82	?
10	26.0	male	no	no	no	no	no	yes	yes	no	no	no	no	0.5	135.0	29.0	3.6	?
11	35.0	female	yes	no	yes	no	no	yes	no	no	no	no	no	0.9	58.0	92.0	3.8	?
12	37.0	female	yes	no	yes	no	no	yes	no	no	yes	no	no	0.6	67.0	28.0	4.3	?
13	36.0	female	no	no	no	no	no	yes	no	yes	no	no	no	0.8	85.0	44.0	4.2	?
14	37.0	female	no	no	yes	yes	yes	yes	no	no	no	no	no	0.8	92.0	59.0	4.2	?
15	67.0	female	yes	no	yes	yes	yes	yes	no	no	yes	yes	no	4.1	105.33	48.0	3.82	?
16	74.0	female	no	no	yes	yes	no	no	yes	no	yes	no	no	2.8	127.0	182.0	2.8	?
17	51.0	female	no	no	yes	yes	yes	yes	yes	no	no	no	no	1.0	105.33	45.0	3.82	?
18	53.0	female	yes	no	no	no	no	yes	no	no	no	no	no	1.5	100.0	100.0	4.0	?
19	32.0	female	no	yes	yes	yes	no	yes	no	no	no	no	no	1.0	66.0	45.0	5.3	?
20	58.0	female	yes	no	yes	no	no	no	yes	yes	yes	no	no	2.0	167.0	242.0	4.1	?
21	44.0	female	no	no	yes	yes	no	yes	no	yes	no	no	yes	0.6	135.0	56.0	3.3	?
22	30.0	female	yes	no	yes	yes	yes	yes	yes	no	yes	yes	yes	2.5	186.0	64.0	3.82	?
23	38.0	female	no	no	yes	yes	yes	yes	yes	no	yes	yes	yes	1.2	118.0	16.0	2.8	?
24	38.0	female	no	no	yes	yes	yes	no	yes	no	no	no	no	0.6	78.0	18.0	4.4	?
25	50.0	male	no	no	yes	no	no	no	yes	yes	yes	no	no	0.9	230.0	117.0	3.4	?
26	42.0	female	no	no	yes	yes	yes	yes	no	yes	yes	no	yes	4.5	105.33	55.0	3.3	?
27	33.0	female	yes	no	no	no	no	no	no	no	no	no	no	1.0	105.33	60.0	4.0	?
28	52.0	female	no	no	no	no	no	yes	no	no	no	no	no	1.5	105.33	89.0	2.9	?
29	58.0	female	no	no	yes	yes	no	no	yes	yes	yes	no	no	1.5	107.0	157.0	3.8	?
30	40.0	female	no	yes	yes	yes	yes	no	yes	no	no	no	no	0.6	40.0	66.0	4.2	?
31	30.0	female	no	no	yes	yes	no	yes	yes	no	yes	no	no	0.8	147.0	128.0	3.9	?
32	44.0	female	no	no	yes	yes	no	no	yes	no	yes	no	no	3.0	114.0	65.0	3.5	?
33	47.0	female	yes	no	no	no	no	yes	no	no	yes	no	yes	2.0	84.0	23.0	4.2	?

Fig. 6.3.17 Data set NN

6.3.6 At Nearest Neighbour Method

i) Large training dataset (I_n, P_n).

ii) Following test point I will be classify.

iii) Calculate distance between other test point in I.

iv) Find minimum distance in other "C".

v) Classification I for test point to current nearest training point "C" and mark it.

vi) All the training data I, are visited, then terminate.

vii) Go to step 2, optimize.

6.3.6.1 Advantage of Nearest Neighbour

i) Easy and simple classification process.

ii) It is demonstrate the nearest neighbour.

iii) Calculate the test static and determines whether the distribution is randomly.

iv) Same class label with a high probability.

6.3.6.2 Dis-Advantage of Nearest Neighbour

i) Very high cost to calculate distances.

ii) Don't use optimization solution.

iii) It is not possible to linear technique.

6.3.7 Case Based Reasoning

Case Based Reasoning (CBR) is the process developed new solution of unsolved problem based on the similar past problems, basically this is human reasoning and thinking but based on his knowledge.

Case : Particular set information.

Based : Knowledge about

Reasoning : Intelligence or new technique.

Possible all Information + Based On Problem Related = Follow Intelligence or New Technique.

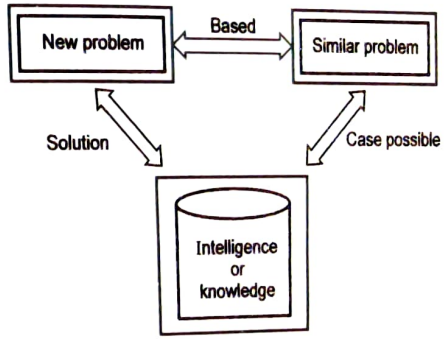


Fig. 6.3.18 CBR

Example :

Problem	Solution
• Key board is not working • Computer P_C : P_4 • Company product : IBM • CPU (P) : Intel • Mouse : OK • Moniter : OK	i) Check key board wire is repair. ii) **Alternative solution** : Check key baord and replace.

Problem	Solution
• Key board is not working • Computer P_C : P_4 • Company product : IBM • CPU (P) : Intel • Mouse : OK • Moniter : OK	i) Check key board wire is repair. ii) **Alternative solution** : Check key baord and replace.

Problem	Solution
• Key board is not working • Computer dual to course • Company product : HCL • CPU (P) : intel • Mouse : OK • Moniter : OK	i) Wire is defected. ii) **Alternative solution** : Change key baord.

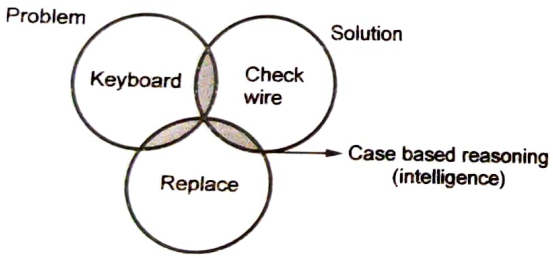


Fig. 6.3.19 Example

6.3.4 CBR Circle

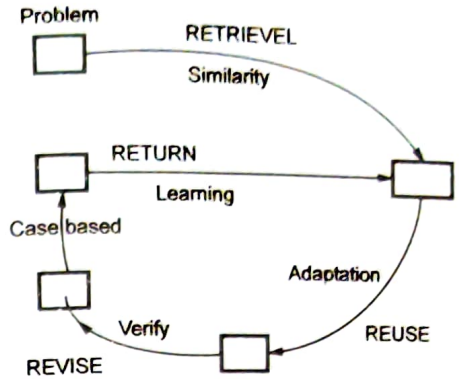


Fig. 6.3.20 CBR circle

i) **Retrieval :**
   i) K-nearest neighbour : Related problem or according similar solution.
   ii) Related problem select best or similar case.
   iii) Easy to solution problem.

ii) **Reuse :**
   i) At a time-different-different approach can adapt according problem.
   ii) Followed different-different case in one problem.

iii) **REVISE :**
   i) Defined similarity between previous solution and new solution of problem
   ii) Check new solution is better, in privious solution.

iv) **RETURN :**
   i) It is a heutistics approach.
   ii) Match to best case of previous problem solution according.

6.3.5 Advantage of CBR

i) Improve performance to solution.
ii) User acceptance is always to follow up new technique.
iii) Low maintenance or case bases are easier to maintain.
iv) High flexibility.
v) Reduce cost.

6.3.7.3 Disadvantage of CBR

i) Problem solution based only knowledge.

ii) Always it does not possible to similar case are priviouse

iii) Case is based on independent problem.

iv) Only depended case are followed.

v) CBR problem solving involves different-different phase like, Retrive, Reuse, Revise, Retarin.

6.3.8 Rough Set Approach

The theory of rough set is a mathematical tool for extracting knowledge from uncertain and incomplete data based information.

The theory of ₹ can be used to find dependence, relationship among data, evaluate the importance of attributes. Discover the pattern of data learn common decision-making rules reduce all redundant objects and attributes, seek the minimum subset of attributes so as to attain satisfying classification. This theory become very popular among scientists around the world and the rough set is now one of the most developing intelligent data analysis. Unlike other intelligent method such as fuzzy set theory, statistical methods, rough set analysis requires no external parameters and uses only information presented in the **Table 6.3.1.** Columns of which are labeled by attribute row by objects of interest and entries of the table are attribute values. Such tables are known as information systems. Attribute value table, data tables. Usually we distinguish in information table are divided in to two disjoint attribute.

i) Decision : Decision table row is define decision rule ("if — then")

ii) Condition : Define decision to condition are satisfied.

For example :

Student record where is mark 100(80-100) and mark 50(40-50) and mark 25(20-30) indifferent different subject. Calculate performance level of related subject according to mark.

Student	Chemistry (S) Mark	Maths (m) Mark	Physic (P) Mark	Level of performance good
1	100	100	25	y
2	50	100	25	n
3	50	100	25	y
4	25	25	25	n
5	50	25	100	n
6	100	25	100	y

Table 6.3.1 Student Record

The table contains six student record concerning here exposed to higher or good pressure endurance test. In the table C_1m and P are condition attribute displaying the percentage or number contact in the record of student in subject (chemistry C) math (M) and physic (P) respectively, where the attribute level of performance good revels the result of the test. The value of condition attribute are as followed (C_1 (80-100)> 3.6 %). 3.5 % = (C, 40-50), (C, (25-30) < 3.5 %.

(M, 80-100) = 0.1 %, (M, 25-30) < 0.1 %, (P, 80-100) = 0.3 %, (P, 25-30) < 0.3 %

Here main problem we are interested how the endurance of the student record base performance depends on the average of C, M and P comprised.

a) If there is a functional dependency between the decision attribute and the condition attribute C, M and P in rough set theory language propound.

i) If the set {2, 4, 5} of all student mark level of performance good after the test the set {1, 3, 6} of poor or no.)

ii) Easily seen that is impossible performance of mark study in subject 2 and 3 display the same feature in terms of attribute C, M and P, but they have different mark of the attribut is not good those information given in Table 6.9.1 is not sufficient to solve our problem.

iii) If the attribute C in the mark 100(80-100) for a certain study record of student then the performance of student in all subject is good, where if the mark of the attribute C is (20-30) 25, then the performance of student is not good. Hence employing attribute C, M and P we can say that student performance t and 6 surely are good. Surely belong to the set {1, 3, 6} where study performance based on mark of subject in student 1, 2, 3 and 6 possible are good, possible belong to the set {1, 3, 6}. Thus the set {1, 6} {1, 2, 3, 6} and {2, 3} are the performance is medium. The upper approximatin and the boundary region of the set {1, 3 6}.

This means that the student performance in mark cannot be determined exactly by the content of mark Chemistry, Maths and Physics in the performance but can be determine only with some following approximation.

I) In fact approximation determine the dependancy (total or partial) between condition and decision attribute.

II) The degree of performance dependancy between condition and decision attribute can be defined as a consistency factor of the decision table. Which is the number of conflicting decision rules to all decision rule we mean rules having the same condition but different decision.

b) **Rules base decision :** The consistency factor for **Table 6.3.1** is 416 = 2/3 hence the degree of dependency between poor and the composition of the performance is 2/3. That means that out of six (60 %) student performance can be mark can be

properly classified as good or not good or not poor on the basis of their subject mark we might be also interested in reducing some of the condition attribute to know whether all condition (mark) are necessary to make write decision specified in a table to this end we will employ/study of the subject mark basic of a reduce of condition attribute. By a reduce we understand a minimal subset mark base condition of attribute which preserves the consistency factor of **Table 6.3.1.**

It is easy to compute that in basic of subject wise mark in **Table 6.3.1** we have to reduce {C, M} and {C, P} intesection of all reduce is called the better performance. In our example the better performance is the attribute C.

C) **Decision of example** : That means that view of the subject mark is most important factor of level of performance poor and connot be eliminated from our consideration where Maths and Physics subject mark can be mutually exchange as factor of poor performance.

6.3.8.1 Process of Rough Set

Suppose we are given two finite, non-empty set U and P, where U is the universe and P is a set of attribute. With every attribute $p' \in P$ we associate a set $V_{p'}$ of its value called the domain of p'. Any subset S of P determine a binary relation I(S) on U which will be called as indiscernibility relation and its is defined as follows.

i) A I(S) B if only if $p'(A) = p'(B) \because p' \in P$ where $p'(A)$ denotes the value of attribute p' for element A.

ii) I(S) is an equivalence relation all equivalence classes of I(B″). Partition determine by B', will be denoted by U/I(S) an equivalence class of I(B″) block of the portion.

iii) If (A, B) belong to I(S) will say that A and B are B-indiscernibile equivalence classes of the relation I(B) are based on B″ elementary set.

iv) Rough set approach elementry set are

$S(A) = \{A \in U : S(A) \subseteq A\}$

$S'(A) = \{A \in U : S(A) \cap A = \phi\}$

6.3.8.2 Application of Rough Set

There are many application of rough set.

i) Medicine : Pharmacology, analysis of relation between the chemical structure and the antimicrobial activity of drugs has been successfully investigated.

ii) Banking : Include evaluation of a bank loan risk and market statige evaluate.

iii) Speech recognition : Obtain in speaker independent speech recognization and acoustics and summerised.

iv) Engineering application : Using vibroacoustics symptoms noise, vibration and process control, machine learning.

v) Linguistics application, environment and databases are other important domains.

6.3.8.3 Advantage of Rough Set

i) The rough set approach seems to be of fundamental importance to AI and connitive science, especially in the areas of machine learning.

ii) Knowledge acquisition, decision analysis, knowledge discovery from database, expert systems, inductive reasoning and pattern recognization.

iii) It seems of particular importance to decision support system and data mining.

iv) Rough set theory is that it does not need any preliminary or additional information about data like (probability statistics and grade of membership fuzzy set theory).

v) The rough set theory has been successfully applied in many real life problems in medical, pharmacology, engineering, banking, financial and market analysis and others.

6.3.8.4 Advance Application of Rough Set

i) Attribute reduction dataset.

ii) Decision rules generation.

iii) Discovery of hidden patterns.

6.4 The Role of Genetic Algorithm

Genetic algorithm has been based on Darwine principal. GA are wide area of experimental and applicable on many related field like Media, Engineering, Insurance, Business, Scientific etc. GA ability to different things for whether a direction is favorable. Genetic will be out standing performance is to be optimize the final result where used real world problems optimization method is that uses population of points at a same time constrast to tradition optimization. Explain in Fig. 6.4.1. Block diagram of Genetic algorithm.

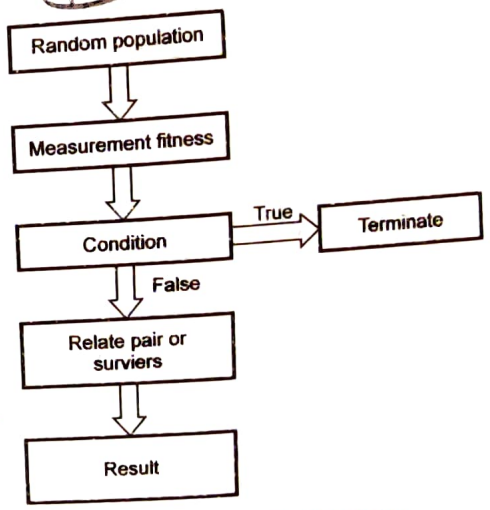


Fig. 6.4.1 Genetic algorithm

We will optimize to GA as a different-different things.

i) GA is based on fitness, parents are selected to reproduce use offspring to finds new generation.

ii) GA followed by component in machine learning.

iii) GA is always optimization creativity.

iv) GA is class based object.

6.4.1 Application of Genetic Algorithm

i) Finance application.

ii) Operation application.

iii) Information system application.

iv) Decision making rule base application.

6.4.2 Implement GA's to Solve the Problem

i) Selection Operator :

i) The higher the selective pressure is more powerful.

ii) Selection operator is unlike the roulette wheel selection.

iii) Improve the fitness of succeeding genes.

iv) Selection operator is two type

a) Rank selection b) Steady-state a selection.

Fitness	Item/8
A	5
B	7
C	1
D	4

⇒

Item/8	Total
A = 5/8	62 %
B = 7/8	87 %
C = 1/8	12 %
D = 4/8	50 %

Selction Fitness = item/8

ii) Crossover Operator :

i) Crossover operator using split parents.

ii) Crossover operator is applied to the mating pool with a hope that it would create a better string.

iii) A cross site is selected at random along the string length and the position values are swapped between two strings.

iv) Crossover is basically three types.

i) 1-point crossover ii) Two-point crossover

iii) Multipoint cross point

Example :

a) One-point crossover

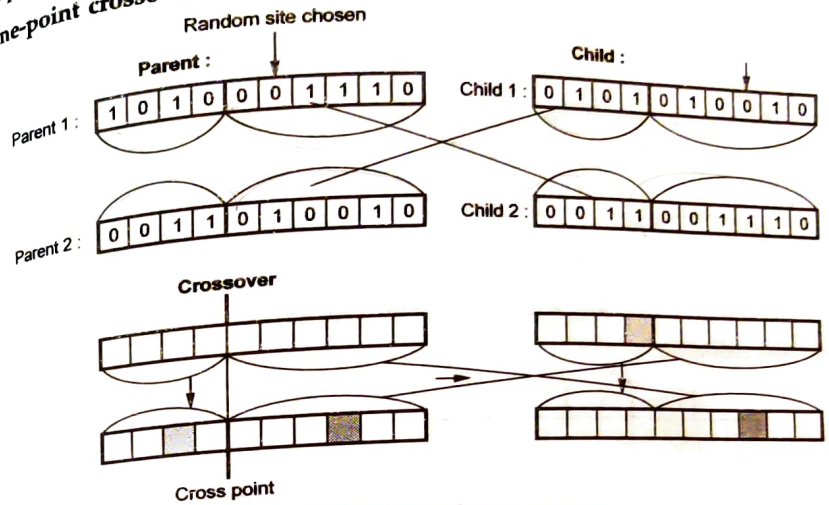


Fig. 6.4.2 One-point crossover

b) Two-point crossover

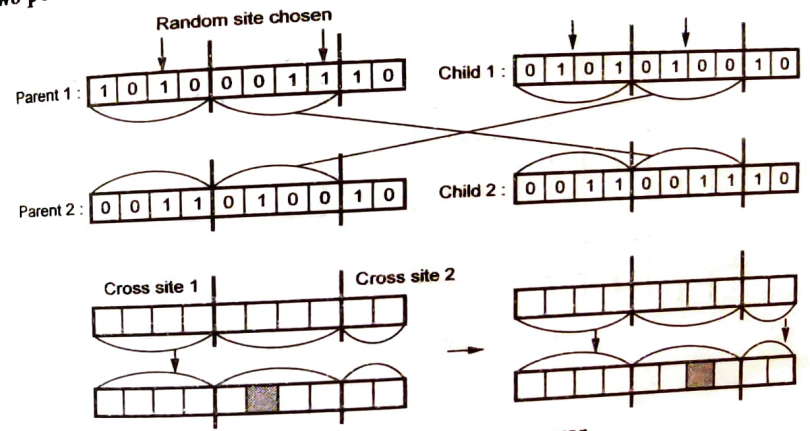


Fig. 6.4.3 Two point crossover

C) Multipoint crossover

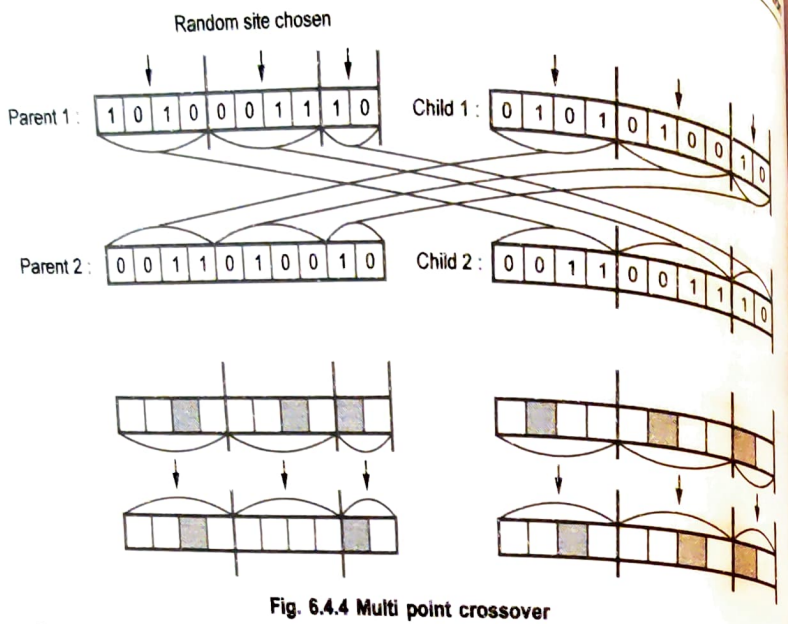


Fig. 6.4.4 Multi point crossover

D) Uniform crossover

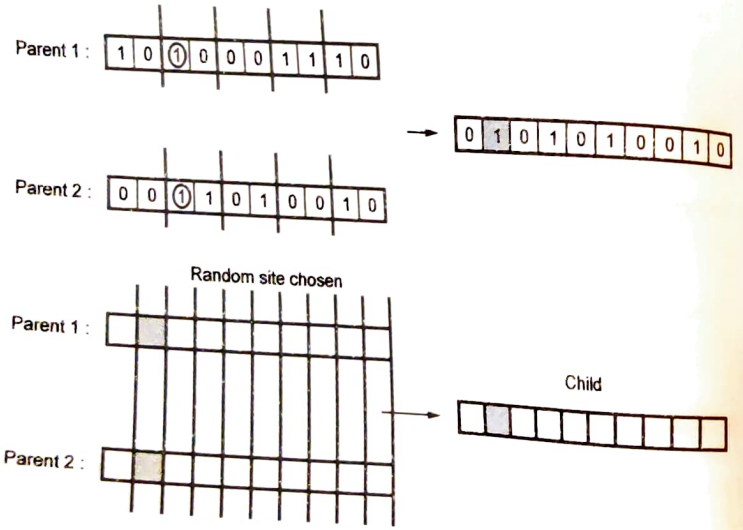


Fig. 6.4.5 Uniform crossover

iii) Mutation :

i) Mutation is maintain diversity of related on population.

ii) It is a parallel process for gene generate.

iii) The bits of the gene are independently mute

iv) Mutain process always generate new genes.

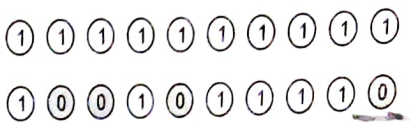


Fig. 6.4.6 Mutation

6.4.3 Advantages of GA

i) Easy to user

ii) Multimodal optimization

iii) Powerful intrachromosomal duplication

iv) GA use to inheritance operatiors.

v) GA to improve population fitness over succeeding generations.

6.4.4 Issue of GA

Following issue are important in GA

i) Representation using string

ii) Fitness function

iii) Reproduction

a) Crossover

b) Mutation

iv) Survivor selection

v) Improve efficiency and increase scability.

6.5 Role of Fuzzy Logic

Fuzzy logic was initiated in 1965, by Lotfi A. Zadeh fuzzy logic started as the language in related argument and persuasion and basically fuzzy logic is used to decision purpose in two-valued logic a proposition like true/false, yes/no, high/low and 0/1 notation. A mathematical membership function μ_x^A is a association with a fuzzy set.

Where is mapping $\mu_{\bar{x}}(A) : A \rightarrow 0.1$

For example :

We could think of the refrigerator cooling speed (CS) of a freezer. Where assume that we want to divide three types of cooling speed min cold, normal cold and max coldest,

but there stage based on crisp set. Freezed cooling speed between 1 to 7 there min cold in winter 1 to 3, freezer 3 to 5 normal in mansoon and max coldest 5 to 7 is summer. The problem that arises is even through a freezer with 3.90 cooling speed is normal cold. It would be classified significantly cooling minimum basically this is called broder problem. Fuzzy set can help overcome this problem solution, it is not 0 and 1. Freez Freezer speed according fuzzy member of more than different-different and therefore give a more realistic view depend on according wheather cooling speed. Following Fig. 6.5.1 based on cooling freezer system.

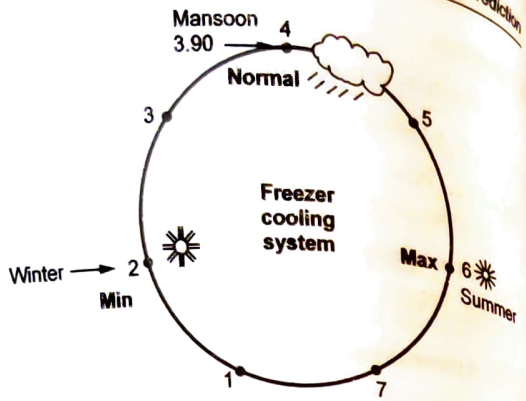


Fig. 6.5.1 Cooling speed

Fuzzy based on Crisp Propertly

The following properties or modifier in thus an operation on fuzzyset.

Name	Operation
i) Commutativity :	$: x \cup y = y \cup x$
	$: x \cap y = y \cap x$
ii) Associative :	$: (x \cup y) \cup z = x \cup (y \cup z)$
	$: (x \cap y) \cap z = x \cap (y \cap z)$
iii) Distributive :	$: x \cap (y \cup z) = (x \cap y) \cup (x \cap z)$
	$: x \cup (y \cap z) = (x \cup y) \cup (x \cup z)$
iv) Demorgan :	$: \overline{x \cap y} = \bar{x} \cup \bar{y}$
	$: \overline{x \cup y} = \bar{x} \cap \bar{y}$
v) Absorption :	$: -(x \cap y) \cup x = A$
	$: (x \cup y) \cap x = A$
vi) Idempotence :	$: x \cup x = A$
	$: x \cap x = A$
vii) Exclusion not satisfied	$: x \cup \bar{x} \neq 1$
	$: x \cap \bar{x} \neq 0$
viii) Transivity :	$: x \subseteq y, y \subseteq z$ then $x \subseteq z$
ix) Involution :	$: \bar{\bar{x}} = x$
x) Contradiction :	$: x \cap \bar{x} = \phi$

6.5.1 Fuzzy Sets Operation

Fuzzy set creates a new set from one or several given set. For example set x and y operation on fuzzy set are as follows.

i) **UNION ($\cup$) :**

i) The union of fuzzy set $\bar{x}$ and $\bar{y}$. There $\bar{x} \cup \bar{y}$.

ii) Then A is membership function of x and y.

iii) μ is a respective membership function.

iv) $\mu_{\bar{x} \cup \bar{y}}(A) = \max(\mu_x(A), \mu_y(A)) = \mu_{\bar{x}}(A) \cup \mu_{\bar{y}}(A)$ $\because A \in x \cup y$

ii) **INTERSECTION ($\cap$) :**

i) The itersection of fuzzy set $\bar{x}$ and $\bar{y}$. There $\bar{x} \cap \bar{y}$.

ii) Then A is membership function of x and y.

iii) μ is a respective membership function.

iv) $\mu_{\bar{x} \cap \bar{y}}(A) = \min(\mu_x A, \mu_y A) = \mu_{\bar{x}}(A) \cap \mu_{\bar{y}}(A)$ $\because A \in x \cap y$

iii) **COMPLEMENT (~ r –) :**

i) It is opposite of for x to $\bar{x}$ and $\bar{x}$ to x.

ii) Then A is membership function of x.

iii) μ is a respective membership function.

iv) $\mu_{\bar{\bar{x}}}(A) = 1 - \mu_{\bar{x}}(A)$ for $A \in x$ or $\bar{x}$

6.5.2 Crisp Logic Operater

Sentences are logic with the words, and, or, if then and if and only if. These are called connectives.

i) A sentence which is operater use by the word **"not"** is called negative.

ii) The other word **"and"** is used to **join** two sentences.

iii) Two sentence formed by connecting two sentences with the word **"or"**.

iv) Two sentence we may construct one of the form "if ... then" it is a conditional statement.

v) The sentence are following "if" is the antecedent and the sentence following "then" is the consequent.

Symbol	Connective	Use
$\neg$	not	$\bar{R}$ or $\neg R$
$\wedge$	and	$R \wedge S$

∨	or	R ∨ S
⇒	if ... then	R ⇒ S
⇔	"if and only if"	R ⇔ S
=	equal	R = S

Using logical formula for the given set of truth values.

R	S	R ∨ S	R ∧ S	– S	R ⇒ S	– (R ∧ S)	R = S
0	0	0	1	1	1	0	1
0	1	1	0	0	0	1	0
1	0	1	0	1	0	1	1
1	1	1	0	0	1	1	1

6.5.3 Use of Fuzzy Logic

i) Railway tracking system
ii) Temperature control
iii) Medical diagnoses system
iv) Antilock braking system
v) Rebotic control
vi) Language processing
vii) Traffic light control

6.5.4 Fuzzy Representation of Freezer Cooling System

Min. cold (A) = {1, if 0 <= Cooling speed (A) < 02 then
0 otherwise}

Normal (A) = {1, if 02 < Cooling speed (A) <= 4.5 then
0 otherwise}

Max cold (A) = {1, if Cooling speed (A) > 4.5 then
0 otherwise}

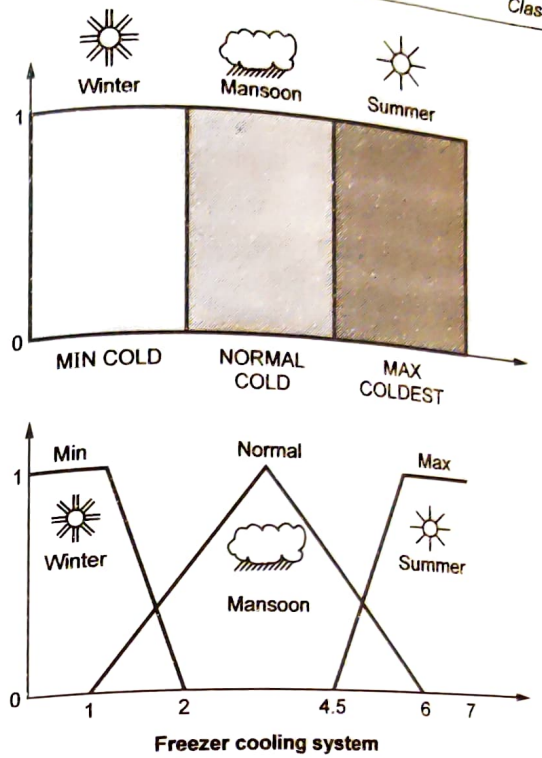


Fig. 6.5.2 Freezer cooling system

Rules for Freezer Cooling System :

Min (A) = {1, if cooling speed (A) <= 01 (02 – cooling speed/10, if
01 < cooling speed (A) < 02, then
0, if cooling speed (A) > 02
}

Normal (A) = {0, if cooling speed (A) <= 01, (1 – (3.5 – cooling speed))
//1.5, if 01 < cooling speed (A) < 3.5,
6 – cooling speed (A)/1.5. Then 3.5 < cooling speed (A) <= 6
}

Max (A) = {0, if cooling speed (A) <= 04 (1 – (7 – cooling speed))/10,
if 4.5 < cooling (A) speed (A) <= 7, Then
1, if cooling speed > 7
}

6.6 Prediction Method Regression

Regression is most common statical mathematical data mining prediction based tool. Regression model can use to one variable with a dataset function as same the target values of another equation that can be used to predict the difference between the predict value and actual expected values.

There are three type of prediction regression.

a) Simple regression/linear regration

b) Multiple linear regration

c) Non-linear regression

   i) Logistic regression

   ii) Exponential regression

   iii) Polynomial regression

6.6.1 Linear Regression

A linear regression technique model can be used to be define measurement relationship between two variable can be observed with a straight line. Data set using two variable A, B where one variable (A) value is called exposition and another variable (B) value is also depend on (A) value exposition variable than it is called dependent variable (B). For following example explain linear regression.

Step 1 : Data

A exposition value : 2 3 4 6 7 8 14 30 45

B dependent value : 2 5 9 10 24 30 48 51 59

Step 2 : Calculate linear regression using this formula

i) Slop (b) ((NTT A·B – (T·TA) (T TB))/(N TTA·A – TTA·A))

ii) Interception (a) = ((TTB·b (T·TA)) / N

iii) F = aA + b

Step 3 : Calculate A, B, A·B, A·A, B·B and total value.

A	B	A·B	A·A	B·B
2	2	4	4	4
3	5	15	9	25
4	9	36	16	81

6	10	60	36	100
7	24	168	49	576
8	30	240	64	900
14	48	672	196	2304
30	51	1530	900	2601
45	59	2655	2025	3481
TT = 119	TT = 238	TT = 5380	TT = 3299	TT = 10072

where N = 8

Step 4 : Calculate slop (b) by using formula;

Slop (b) = ((N TT A·B – (TTA) (TTB)) / (NTTA^2 – TTA^2)

= (8 * (5380) – (119) (238)) / (8× 3299 – 3299)

= (43040 – 28322)/(26392 – 3299) = 14718/23093

b = 0.637

Step 5 : Calculate Intercept (a) = (TTB – b (TTA))/N

Intercept a = ((238 – (0.637) * (119))/8 = (238 – 75.803)/8 = 162.197/8

a = 20.27

Step 6 : Calculate linear regression

f = aA + b

f = aA + b

f = 20.27 * 119 + 0.637

f = 2412.13 + 0.637

f = 2412.76

Step 7 : F = aA + b graph the linear regression

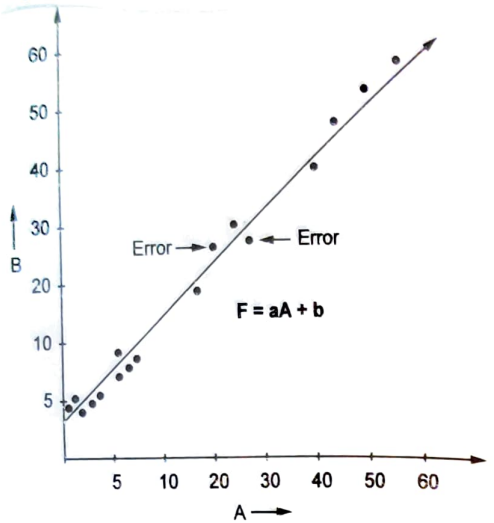


Fig. 6.6.1

6.6.2 Multiple Linear Regression

Multiple linear regression based on linear regression difference only pridictors can be use more then two straight line but not display same dimension using pixel for example previous example based graph.

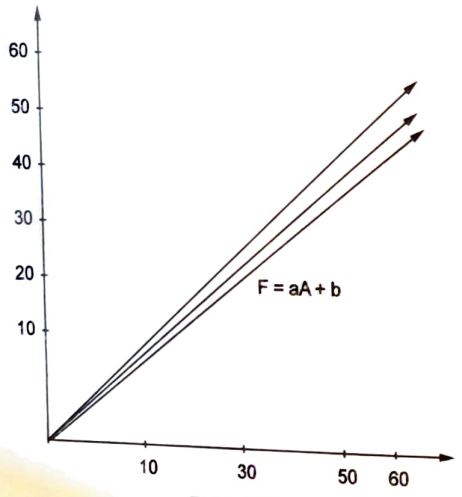


Fig. 6.6.2

Example 2 : Linear regression :

A : 3 7 8 14 21 24

B : 24 30 45 50 59 61

Solve : N = 6

A	B	A·B	A^2
3	24	72	9
7	30	210	49
8	45	360	64
14	50	700	196
21	59	1239	441
24	61	1464	576
TT = 77	TT = 259	TT = 4050	TT = 1335

Linear equation aA + b

$$f = aA + b$$

But linear equation calculate slop and interception prediction

$$f = a_0 + a_1 N$$

$$f = a_0 + a_1 A_1 + a_2 A_2 + a_3 A_3 + a_4 A_4 + a_5 A_5 + a_6 A_6$$

where : $\bar{A} = \frac{A}{N}$ $\bar{B} = \frac{B}{N}$

$\bar{A} = 77/6 = 12.8$ $\bar{B} = \frac{259}{6} = 43.1$

$$a_1 = \frac{(A_1 - \bar{A})(B_1 - \bar{B}) + (A_2 - \bar{A})(B_2 - \bar{B}) + (A_3 - \bar{A})(B_3 - \bar{B}) + (A_4 - \bar{A})(B_4 - \bar{B}) + (A_5 - \bar{A})(B_5 - \bar{B}) + (A_6 - \bar{A})(B_6 - \bar{B})}{(A_1 - \bar{A})^2 + (A_2 - \bar{A})^2 + (A_3 - \bar{A})^2 + (A_4 - \bar{A})^2 + (A_5 - \bar{A})^2 + (A_6 - \bar{A})^2}$$

$$a_1 = \frac{(3 - 12.8)(24 - 43.1) + (7 - 12.8)(30 - 43.1) + (8 - 12.8)(45 - 43.1) + (14 - 12.8)(50 - 43.1) + (21 - 12.8)(59 - 43.1) + (24 - 12.8)(61 - 43.1)}{(3 - 12.8)^2 + (7 - 12.8)^2 + (8 - 12.8)^2 + (14 - 12.8)^2 + (21 - 12.8)^2 + (24 - 12.8)^2}$$

$$a_1 = \frac{(-9.8)(-19.1)+(-5.8)(-13.1)+(-4.8)(1.9)+(1.2)(6.9)+(8.2)(15.9)+(11.2)(17.9)}{96.4+33.64+23.04+1.44+67.24+125.44}$$

$$a_1 = \frac{187.18+75.98+(-9.12)+8.28+130.38+200.38}{347.56}$$

$$a_1 = \frac{187.18+66.86+8.28+130.38+200.48}{347.56}$$

$$a_1 = 593.18/347.56$$

$$a_1 = 1.70$$

$$\boxed{a_1 = 1.70}$$

where

$$a_0 = \bar{B} - A \cdot a$$

$$a_0 = \bar{B} - Aa_1$$

$$a_0 = 43.1 - (12.8)(1.70)$$

$$a_0 = 43.1 - 21.83$$

$$\boxed{a_0 = 21.27}$$

$$F = a_0 + a_1 N$$

$$F = 21.27 + 1.70 \times 6$$

$$F = 21.17 + 10.2$$

$$F = 31.47$$

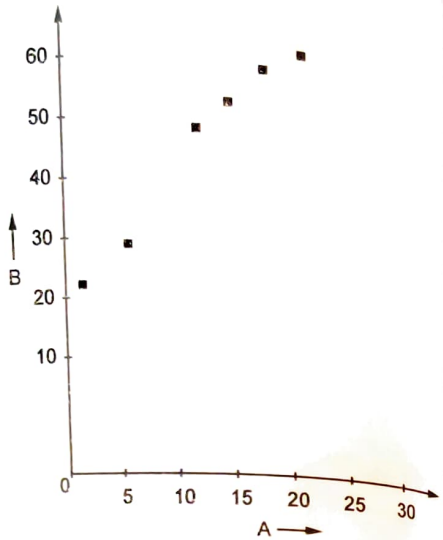


Fig. 6.6.3

6.6.3 Non-linear Regression

The basic idea of non-linear regression is the same as that of linear regression if linear regression is define relation between two variable or value of variable to straight line, but non-linear process is itrative the data could be preprocessed to build or make to stratist linear regression following model of non-linear regression are as follows

i) Exposition regression model
ii) Power regression model
iii) Content increment regression model
iv) Polynomial regression model.

i) Exposition regression model : Where exponential model is $(A_1, B_1)(A_2, B_2) \ldots (A_n B_n)$ where exponential $f = ae^{bA}$.

ii) Power regression model : Where power model is $(A_1, B_1)(A_2, B_2) \ldots (A_n B_n)$ where regression is $f = aA^b$.

iii) Content increment regration model : Where content increment regration model is $(A_1 B_1) \ldots (A_n B_n)$ where content regration $f = \frac{aA}{b+A}$

iv) Polynomial regration model :

$$F = a_0 + a_1 \bar{A} + \ldots + a_n A^n$$

Graph of non-linear regration : Where $(A_1, B_1)(A_2, B_2) \ldots (A_n, B_n)$, $f = y(A)$ where y (A) is non-linear of x.

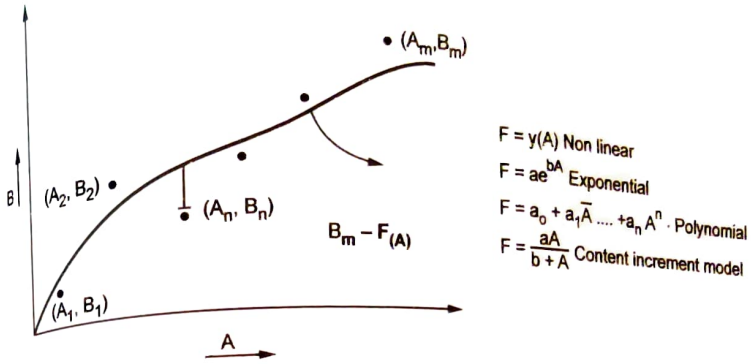


Fig. 6.6.4

a) Using non-linear regration calculate difference r where

$$r = \sum_{n=1}^{n} (f_n - ae^{bAn})^2$$ then calculate a and b

Calculate a :

$$a = \frac{\partial\, TTr}{\partial_a} = \sum_{n=1}^{n} 2(f_n - ae^{bA_n})(-e^{bA_n}) = 0$$ then

Calculate b :

$$b = \frac{\partial\, TTr}{\partial_b} = \sum_{n=1}^{n} 2(f_n - ae^{bA_n})(-aA_n e^{bA_n}) = 0$$

6.6.4 Logistic Regression

Logistic regression is statistical method is similar to the one for multiple regression method. Logistic regression is depend binary variable (value) where is 1 or true and 0 or false.

Purpose of logistic regression is to find best relationship between 1 and 0, based on its characteristic (dependent variable and response variable). First identify depend variable it is also binary variable calculate following in logistic regration.

i) Logistic L or $P = \frac{L}{1-L}$ where L or P = probability

$$\text{Logistic L or P} = \frac{\text{probability of present value}}{\text{probability of absent value}}$$

$$\text{L or P} = \log\left(\frac{L}{1-L}\right)$$

ii) Logistic true or false response :

where A = 0 (false), A = 1 (True)

iii) Regression coefficient :

Logistic L or $P = a_0 + a_1A_1 + a_2A_2 + \ldots + a_nA_n$

where $a_0, a_1 \ldots a_n$ is coefficient

iv) Ration (calculation exposition)

Logistic L or $P = \frac{L}{1-L}$ Exposition (e) is logistic

Logistic $L = e^{a0} * e^{a1A1} * e^{a2A2} * \ldots * e^{anAn}$

v) Interpretation

$$\text{Logistic L} = \frac{1}{1 + e^{\text{logistic or log (L)}}}$$

Example of logistic using regration tool

Where p is the probability that y = 1 and $X_1, X_2, \ldots, X_n$ are the independent variable (predictors). $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are known as the regression coefficients, which have to be estimated from the data. Logistic regression estimates the probability of a certain event occurring. Logistic regression thus forms a predictor variable (log(p/(1-p) which is a linear combination of the explanatory variables. The values of this predictor variable are then transformed into probabilities by a logistic by a logistic function. Such a function has the shape of an S. On the horizontal axis we have the values of the predictor variable and on the vertical axis we have the probabilities.

Dataset description

Age (Years)
Sex (1 - male, 0-female)
Weight in KG

Height in CM.
Diet (1- veg., 0- nonveg.)
BPL(Blow Poverty Line) (1-yes,0-no)
Livingzone (1 for Rural,0 f0r Urban_)
Fedu (Father education)

Data set

age	sex	weight	height	diet	bpl	livingzoe	edu	fedu	medu	addic	dies
42	1	62	170	1	1	1	3	0	0	2	1
40	1	55	16	1	1	0	1	0	0	3	1
65	0	46	150	1	1	0	0	0	0	2	2
70	1	56	162	0	1	0	0	0	0	3	2
45	0	52	150	0	1	0	1	0	0	2	2
50	1	55	162	0	1	0	0	0	0	2	2
36	1	60	161	0	1	0	1	0	0	2	1
66	0	52	150	0	0	0	0	0	0	2	1
60	1	50	65	0	1	0	2	0	0	2	2
30	1	50	160	0	0	0	2	1	0	2	2
90	1	60	170	1	1	0	2	0	0	3	2
32	0	42	160	1	1	1	1	0	0	0	2
48	0	51	160	1	1	1	3	0	0	2	1
73	0	48	158	1	1	1	1	0	0	0	1
35	0	60	160	1	1	1	2	2	0	0	1
30	0	45	161	1	0	1	4	3	1	0	1
30	1	50	160	0	1	0	2	1	3	3	2
26	1	49	163	0	1	0	4	0	0	2	2
28	1	56	162	0	1	0	1	0	0	2	2
40	1	50	164	1	0	0	0	0	0	2	2

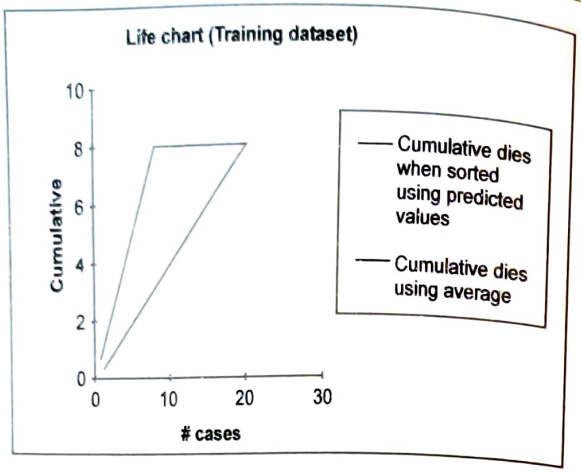


Fig. 6.6.5 Result

Decile	Mean	Std. Dev.	Min.	Max.
1	1	0	0	1
2	1	0	0	1
3	1	0	0	1
4	1	0	0	1
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0

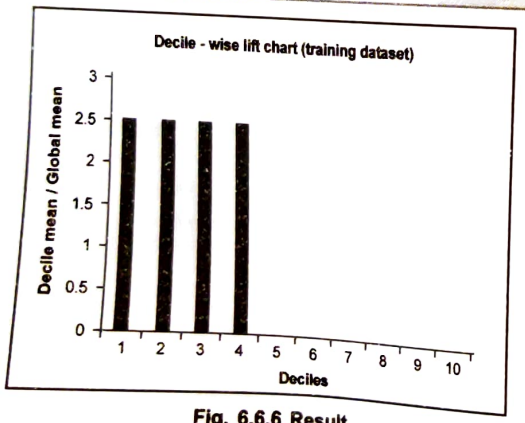


Fig. 6.6.6 Result

Data	
Trainig data used for building the model	['surveydata12.csv']'surveydata12'!$A2:$L$21
# Records in the training data	20
New data	['surveydata12.csv']'surveydata12'!$A2:$L$21
# Records in the new data	20

Variables											
#Input Variables	11										
Input variable	age	sex	weight	hight	diet	bpl	livinzoe	edu	fedu	medu	addic
Output variable	dies										
Constant term present	yes										

Parameters/Options	
# Iterations	50
Marquardt overshoot factor	1
Initial cutoff probability value	0.5
Confidence Level %	95
Perform subset selection	Yes
Subset selection producer	Forward selection
Maximum size of best subsets	11
# Best subset	1
Show coviance matrix	Yes

Output options chosen
Summary report of scoring on training data
Detailed report of scoring on training data
Lift charts on training data
Detailed report of scoring on new data

occurrences in training data

Prob.

0.4 <- Success Class

0.6

	P-value	Odds	95 % Confidence Interval	
	0.34995615	3.73391E-34	2.09748E-34	5.37034E-34
	0.92724645	0.96620131	0.461896	2.0211506
	0.80681831	0.00293118	0	5.87572E+17
	0.29849535	8.89869499	0.14443369	548.256897
	0.49573123	0.86956406	0.58165348	1.2999866
	0.87128133	0.01556252	0	1.15099E + 20
	0.70029587	0.0000007	0	1.54089E + 25
	0.68880928	3355023000	0	*
	0.94628119	0.42322668	0	30853100000
	0.90757787	0.03451739	0	1.68099E + 23
	0.85207957	189.3627014	0	1.63764E + 26
	0.87774616	0.06968283	0	3.83064E + 13

Residual df	8
Residual Dev.	0.0923662
% Success in traning data	40
# Iterations used	11
Multiple R-squared	0..99656892

	age	sex	weight	hight	diet	bpl	livingzoe	edu	fedu	medu	addic
					91.0381546	55.37468066	-1266.81995	106.6438751	-205.720154	-201.344177	34.9043541
					-3.41758704	0.27998081	3.94176245	-0.859366	1.83111715	-0.3646883	0.5302699
					-107.788452	26.66061401	543.2032471	-29.4125671	51.07870102	-64.1368103	72.84882355
					-1.65765572	-34.2516556	13.10546207	1.19355512	-17.245533	18.69153404	-12.8688116
					2.63317609	-3.68995497	0.5336079	0.67503422	-1.10605085	0.27699783	-201.692078
					660.0741577	264.2183228	-1077.49255	106.8001328	-41.4596939	16.79881287	-392.574615
					264.2183228	1355.894897	-457.688232	-84.0429001		-392.574615	101.614212

	-1266.81995	3.94176245	543.2032471	13.10546207	3.68993497	-1077.49255	-457.688232	2999.726318	-403.351715	288.5884094	-107.511406	641.0836182
	106.6438751	-0.859366	-29.4125671	1.19355512	0.5336079	106.8001328	-84.0429001	-403.351715	162.8593903	-218.111572	68.51412964	-138.443359
	-205.720154	1.83111715	51.07870102	-17.245533	0.67503422	-41.4596939	716.1317749	288.5884094	-218.111572	841.1827393	-513.678162	238.8925934
	-201.344177	-0.3646883	-64.1386103	18.69153404	1.10605085	16.79881287	-392.574615	-107.511406	68.51412964	-513.678162	790.8094482	-151.730576
	34.9043541	0.5302699	72.84882355	-12.8688116	0.27699783	-201.692078	101.614212	641.0836182	-138.443359	238.8925934	-151.730576	299.8739319

Actuals	RSS	Cp	Probability	Model (Constant present in all models)											
				1	2	3	4	5	6	7	8	9	10	11	12
1	8.09257701	-6.75100422	0.99904859	Constant	weight	*	*	*	*	*	*	*	*	*	*
2	7.52318527	-5.40206194	0.99987024	Constant	weight	livingzoe	*	*	*	*	*	*	*	*	*
3	7.33388472	-3.61871702	0.99991482	Constant	weight	hight	livingzoe	*	*	*	*	*	*	*	*
4	7.21716928	-1.75180602	0.99991554	Constant	weight	hight	bpl	livingzoe	*	*	*	*	*	*	*
5	7.13809729	0.15782592	0.99988931	Constant	weight	hight	diet	bpl	livingzoe	*	*	*	*	*	*
6	7.0896349	2.10244036	0.99977911	Constant	weight	hight	diet	bpl	livindzoe	edu	*	*	*	*	*
7	7.05192327	4.05934143	0.9994998	Constant	sex	weight	hight	diet	bpl	livindzoe	edu	medu	*	*	*
8	7.03099844	6.03439856	0.99821407	Constant	sex	weight	hight	diet	bpl	livindzoe	edu	medu	addic	*	*
9	7.01799428	8.02046204	0.98983127	Constant	sex	weight	hight	diet	bpl	livindzoe	edu	fedu	medu	addic	*
10	7.00792313	10.00905514	0.92627275	Constant	sex	weight	hight	diet	bpl	livindzoe	edu	fedu	medu	addic	
11	6.99999952	12	1	Constant	age	sex	weight	hight	diet	bpl	livindzoe	edu	fedu	medu	addic
12															

Cut off Prob.Val. for Success (Update table) 0.5 (Updating the value here will NOT update value in summary report)

Classification Confusion Matrix

Actual Class	Predicted Class 1	Predicted Class 2
1	8	0
2	0	12

Error Report

Class	# Cases	# Errors	% Error
1	8	0	0.00
2	12	0	0.00
Overall	20	0	0.00

Overall (secs)	3.00

XL Miner : Logistic Regression-Classification of Training Data

Data Range ['surveydata 12.csv']'surveydata 12'!$A$2:$L$21 **Back to Navigator**

Cut off Prob.Val. for Success (Updata table) 0.5 (Updating the value here will NOT update value in summary report)

Row Id.	Predicted Class	Actual Class	Prob. for 1 (success)	Logs odds	age	sex	weight	height	diet	bpl	livingzoe	edu	fedu	medu	addic
1	1	1	1	23.2093916	42	1	62	170	1	1	1	3	0	0	
2	1	1	0.992671714	6.62253656	40	1	55	16	1	1	0	1	0	0	2
3	2	2	7.73674E-11	-23.2824562	65	0	46	150	1	1	0	0	0	0	3
4	2	2	0.000497336	-7.66574809	70	1	56	162	0	1	0	0	0	0	2
5	2	2	0.002073739	-6.17632605	45	0	52	150	0	1	0	1	0	0	3
6	2	2	0.00159356	-6.44019016	50	1	55	162	0	1	0	0	0	0	2
7	1	1	0.985944841	4.25061092	36	1	60	161	0	1	0	1	0	0	2
8	1	1	0.999707053	8.13522624	66	0	52	150	0	0	0	0	0	0	2
9	2	2	0.00279757	-5.87620265	60	1	50	65	0	1	0	2	0	0	2
10	2	2	0.000665196	-7.31476317	30	1	50	160	0	0	0	2	1	0	2
11	2	2	0.001427387	-6.55048131	90	1	60	170	1	1	0	2	0	0	2
12	2	2	0.002765989	-5.88758707	32	0	42	160	1	1	1	1	0	0	3
13	1	1	0.997950539	6.18812703	48	0	51	160	1	1	1	3	0	0	0
14	1	1	0.997756924	6.09766138	73	0	48	158	1	1	1	1	0	0	2
15	1	1	1	25.76311293	35	0	60	150	1	1	1	2	2	0	0
16	1	1	0.99350153	7.33812368	30	0	45	161	1	0	1	4	3	1	0
17	2	2	0.000220075	-8.42132079	30	1	50	160	0	1	0	2	1	3	3
18	2	2	2.045795-10	-22.3095767	25	1	49	163	0	1	0	4	0	0	2
19	2	2	0.012645772	-4.35770593	28	1	56	162	0	1	0	1	0	0	2
20	2	2	0.00067877	-7.29454317	40	1	50	164	1	0	0	0	0	0	2

Data Source	
Work Book path	F :
Work Book Name	surveydata12.csv
Training Range	[surveydata12]!$A$2:$L21
#Training	20
#Variables in Data set	12
#Selected Variables	12

Data Dictionary

	age	sex	weight	hight	diet	bpl	livingzoe	edu	fedu	medu	addic	dies
Number in Data set												
Variable Type	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous	Continuous	Binary
Variable Data Type	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number

Mining Scheme

	age	sex	weight	hight	diet	bpl	livingzoe	edu	fedu	medu	addic	dies
Selected Variables												
Variable Type	Input	Input	Input	Input	Input	Input	Input	Input	Input	Input	Input	Output
Inputs Normalised	No											
Constant Term Present	Yes											

Classes in Inputs Data set	
# Classes	2
Class 1 (Success)	1
Class 2	2

Prior class probabilities

Class	Prob.
1	0.4
2	0.6

Model

Input Variables	Coefficient
Constant Term	-76.970436
age	-0.0343831
sex	-5.8323507
weight	2.1859045
height	-0.1397633
diet	-4.1628895
bpl	-14.173749
livingzone	21.9337235
edu	-0.8598474

fedu	-3.3662922
medu	5.24366426
addic	-2.6638014

Other Options	
# Iterations	50
Marquardt overshoot factor	1
Initial cutoff probability value	0.5
Confidence Level %	95

6.6.4.1 Applications of Logistic Regration

i) Doctor : For pateint servives treatment disease.

ii) Student : Course guidence

iii) Lone : Banking sector

iv) Insurance company : Customer requirement prediction

6.6.4.2 Advantages of Logistic Regration

i) Logistic is better than result as non-linear regration

ii) Independent variable or value have equal distributed

iii) Easy to calculate interpretation

iv) Minimum cost regration method

6.6.4.3 Disadvantages of Logistic Regration

i) Independent variable have equal distributed it does not require always.

ii) Logistic regration is different-different homogenious assumption.

6.7 Introduction of Tools Such as DB Miner/ WEKA/DTREG Data Mining Tools

Different types of data mining tool are available in the marketplace today, each with their own strengths and weaknesses. Internal auditors need to be aware of the different kinds of data mining tools available and recommend the purchase of a tool that matches the organization's current detective needs or analysis purpose. This should be considered as early as possible in the project's life cycle, perhaps even in the feasibility study. Most data mining tools can be classified into the three categories : a) traditional tool b) dashboard tool c) text - mining tool.

a) **Traditional tools** : Help companies establish patterns and trends by using a number of complex algorithms and techniques.

b) **Dashboards Tools** : Reflect data changes and updates on screen - often in the form of a chart or table.

c) **Text- mining Tools** : Its ability to mine data from different kinds of text from Microsoft Word and Acrobat PDF documents to simple text files.

DB Miner :

The system has been developed for interactive mining of multiple-level knowledge in large relational databases and data warehouses. DB Miner implements a wide spectrum of data mining function, including characterization, comparison, association, classification, prediction and clustering many useful etc. By incorporation of several interesting data mining techniques, including attribute - oriented induction, progressive deepening for mining multiple - level rules and meta-rule guided knowledge mining, including attribute- induction, statistical analysis, the system provides a user - friendly, interactive data mining environment with good performance and easy to understand.

6.7.1 DB Miner : Project Description

- If data mining techniques, including attribute-oriented induction, meta-rule guided knowledge mining and progressive deepening for mining multiple-level rules etc., implements a wide spectrum of data mining functions including generalization, characterization, association, classification and prediction etc.
- It performs interactive data mining at multiple concept levels on any user - specified set of data in a database using an SQL - like Data Mining Query Language, DMQL. Users may interactively set and adjust various thersholds, control a data mining process, perform including generalized relations, generalized feature tables, multiple forms of generalized rules, visual presenation of rules, charts, curves etc.
- The data mining process may utilize user defined set-grouping concept hierarchies which can be specified flexibly, adjusted dynamically based on data distribution and generated automatically for numerical attributes. Efficient implementation techniques explored using different data structures, including generalized relations and multiple - dimensional data cubes.
- Both UNIX and PC (Windows/NT) versions of the system adopt client /server architecture.

Major functional DBMiner

- DBMiner evolution evaluator
- DBMiner deviation evaluator
- DBMiner user interfaces
- DBMiner charaterizer
- DBMiner discriminator
- DBMiner association rule finder
- DBMiner data classifier
- DBMiner predictor
- DBMiner meta-rule guided DBMiner

6.8 Weka Tool

Waikto Environment for knowledge Analysis (Weka) is a collection machine learing algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. **Weka contains tools for data pre-processing,** classification, regression, clustering association rules and visualization only. It is also well- suited for developing new machine learning schemes.

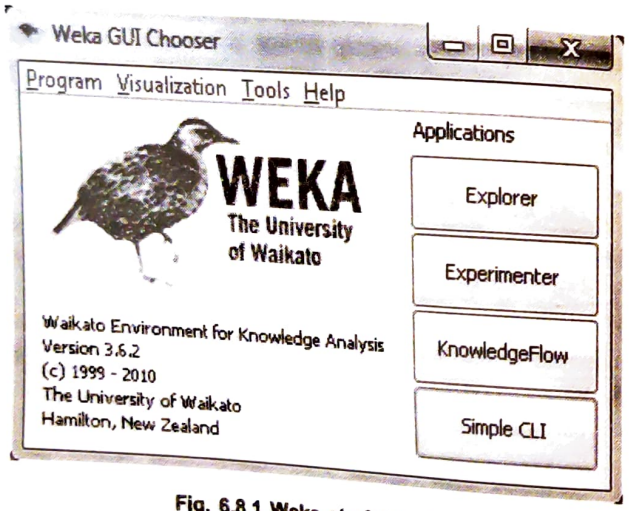


Fig. 6.8.1 Weka starful tool (1)

When you start WEKA, the GUI chooser pops up and lets you choose four ways to work with WEKA and your data. For all the examples in this article series, we will choose only the **Explorer** option.

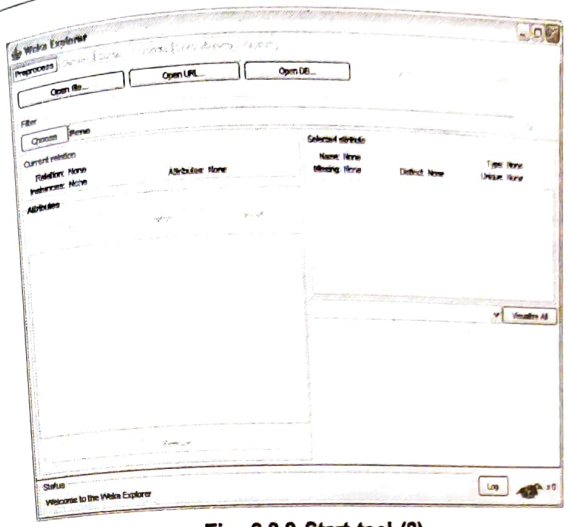


Fig. 6.8.2 Start tool (2)

6.8.1 Data Set for WEKA

To load data into WEKA tool, we have to put it into that will be understood. WEKA's preferred method for loading data is in the Attribute Relation File Format (ARFF) format, where you can define the type of data being loaded, then supply the data itself. In the file, you define each column and what each column contains.

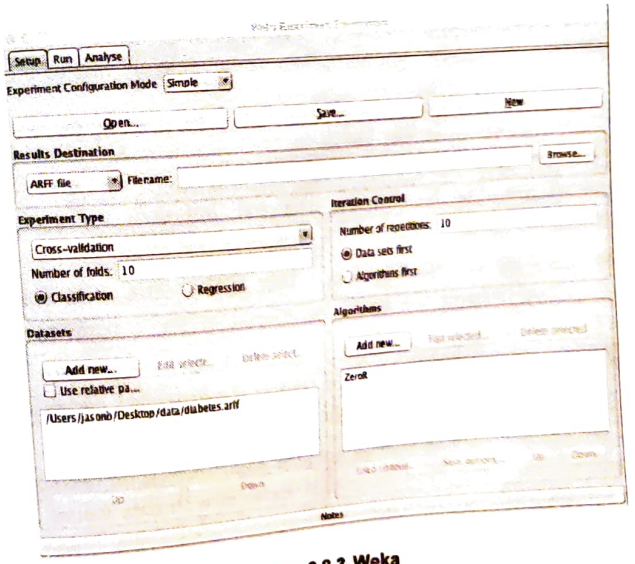


Fig. 6.8.3 Weka

Loading the data into WEKA

Now that the data file has been created, it's time to create our regression model. Start WEKA and then choose the **Explorer.** You'll be taken to Explorer screen, with the **Pre-process** tab selected. Select the **Open file** button and select the ARFF file you created in the section above. After selecting the file.

6.9 DTREG (pronounced D - T- Reg)

This data can be used to create models to make prediction. Many techniques have been developed for predictive modeling, and there is an art to selecting and applying the best method for a particular problem or situation. DTERG implements the most powerful predictive modeling methods that have been developed inducing, Decision Three Forests and Three Boost as well as Neural Networks, Support Vector Machine, Gene Expression Programming and Symbolic Regression, K- Means Clustering, Linear Discriminate Analysis Regression models and Logistic Regression models.

6.9.1 DTREG Features

- **Ease of use.** DTREG is a robust application that is installed easily on any Windows system. DTREG reads Comma Separated Value data files that are easily created from almost and data source.
- **Classification and Regression Tress.** DTREG can build Classification Tress where the target variable being predicted is categorical and Regression Trees where the target variable is continuous according data to our data set like as income or sales volume.
- **Automatic tree pruning.** In this tool uses V- fold cross - validation to determine the optimal tree size. This procedure avoids the problem of "over-fitting" where the generated tree fits the training data well but does not provide accurate predictions of new data.
- **Surrogate variables for missing data :** This allows cases with some available values and some missing values to be utilized to the maximum extent when building the model.
- **Visual display of the tree.** DTREG can display the generated decision tree on the screen; write it to a .jpg or .png disk file formate.
- **DTERG accepts text data as well as numeric data.** There is no need to code them as numeric values. If you have categorical variables with data values such type of "Male", "Female", "Married, etc.

- **DTREG .NET Class Library.** The DTREG .NET Class Library can be called from application programs to generate models and compute predicted target values using a model generated by DTREG.
- **Project files for saving analyses.** DTREG save all of the information about variables, analysis parameters as well as the generated report and tree in project file.
- **Scoring to predict values.** You can use DTREG to score a new dataset and predict values for the target variable.
- **Generated scoring source code.** The "Translate" function in DTREG generates C, C++ and SAS Source code to compute predicted values.
- **Heavy duty capability.** DTREG can build classification trees with predicator variables that have hundreds categories by using an efficient clustering algorithm. Many other decision tree programs limit predictor variables to 16 or categories.

Review Questions

1. *What is classification ? Discuss any method.*
2. *Explain linear regression with suitable example.*
3. *Explain logistic regression method.*
4. *Explain advantages and disadvantages of neural network.*
5. *Explain rough set approach.*

❑❑❑

7 Data Mining for Business Intelligence Application

Syllabus

Data mining for business applications like balanced scorecard, Fraud detection, Clickstream mining, Market segmentation, Retail industry, Telecommunications industry, Banking and finance and CRM etc.,

- *Data analytics life cycle : Introduction to big data business analytics - State of the practice in analytics role of data scientists.*
- *Key roles for successful analytic project - Main phases of life cycle - Developing core deliverables for stakeholders.*

Contents

7.1 Business Intelligence Data Mining

One of the basic factor that influence the powerful process to increasing performance of a business by the intelligent use to make effective and timely decision making.

Business intelligence methodologies are varied and complex have a brode area of application (sorting, analysis) can decision support system. Like customer relationship management, chain optimization, forecasting and assortment optimization.

For example :

Through business intelligence data mining. Use company can identify or maintain profile of customer in company loan data and product data (place strategies and policies) may not be easy to maintain normal reporting. While business intelligence data mining tool using properly easy, timely maintain all reporting of company every year for decision making.

7.2 Business Intelligence Improve Performance in Management

i) Organization control

ii) Maintain co-ordination in organization.

iii) Define strategies that would several benefit both organization and customer satisfaction.

iv) Learn feedback and how progress to performance.

v) Continually assess the performance and create plan.

7.3 Applied in DM in Business Intelligence

i) Text mining

ii) MOLAP

iii) Management system

iv) Decision making system

v) Web mining

vi) Data warehouse

vii) Visualization representation

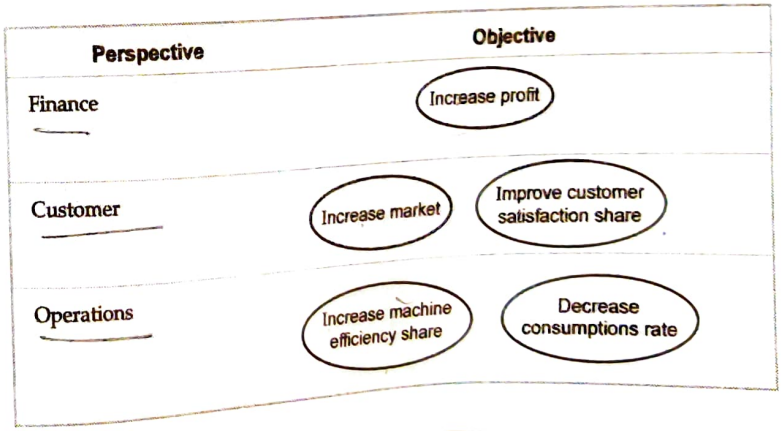
viii) Geographic information system.

7.4 Application of Business Intelligence

7.4.1 Balanced Scorecard

The Blanced Scorecard (BSC) in this application used for measuring success and setting goals from financial and operational viewpoints. With these measures, leaders can manage their strategic vision and adjust it for change. The balanced scorecard links performance measures by looking at a business's strategic vision from four different perspectives: Financial, Customer, Learning and Internal Business Processes. Each of the four perspectives is considered under various parameters. These parameters include goals that have to be achieved in order to become successful. Key performance indicators (KPIs) are parameters that will be used to know if success is achieved; and targets are quantitative value that will be used to determine the success of the KPI. The task to set target is very hard and the result of target-setting in the workplace seems unable to reach satisfaction. Setting targets can be fraught with many problems. For example, if set is too high, and targets create stress and de-motivation; if set is too low, targets encourage complacency. The history values may play important role to set target correctly.

We have data set for a company in Egypt that used BSC framework decade ago. We implement our approach at Corporate Balanced Scorecard as in Fig. 7.4.1. It has more than more observations values for different KPIs. It is used to review and monitor the corporate objectives and KPIs. This BSC contains four perspectives. Finance perspective is the highest priority one that answers question on how the firm looks to shareholders. Customer perspective answers question on how customers see the firm. Operations perspective answers question how well it manages its operational processes. Learning and growth perspective answers question if the firm can continue to improve and create value.



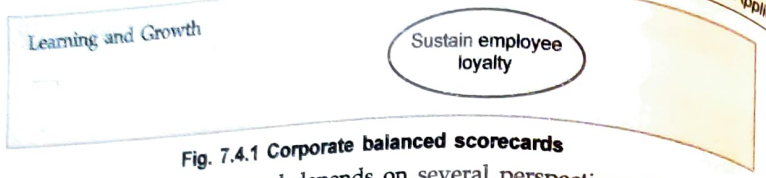


Fig. 7.4.1 Corporate balanced scorecards

The corporate balanced scorecard depends on several perspectives. Every perspective includes several objectives. For example "Finance" perspective includes "Increase Profit Objective". Form customer perspective the objective is to "Increase market share" and "Improve Customer Satisfaction". From operations perspective it is required "Increase Machine Efficiency" and "Decrease Consumptions Rate". Learning and growth perspective includes "Sustain Employee Loyalty" objective.

- **Monthly Shipping "Ton" compared to same month last year**

This KPI is a measure for "Increase profit" objective. It shows how many tons are produced compared to the same month last year.

- **Number of New Customers**

This KPI is a measure for "Increase Market Share" objective. It shows the number of new customers.

- **% of Orders Delivered on-time**

This KPI is a measure for "Improve customer satisfaction" objective. It shows the percentage of orders delivered on-time.

- **% Total Plant Waste**

This KPI is a measure for "Increase machine efficiency" objective. It shows percentage of total plant waste in the manufactory process.

- **Average Material Consumption "Kg/Ton produced"**

This KPI is a measure for "Decrease consumptions rate" objective. It shows average material consumption "Kg/Ton produced".

- **Number of Employees Leaving**

This KPI is a measure for "Learning and growth" objective. It shows the loyalty of employees through measuring the number of employees leaving yearly.

Objective	KPI
Increase profit	Monthly shipping "Ton" compared to same month last year.
Increase market share	Number of new customers.
Improve customer satisfaction	% of orders delivered on-time
Increase machine efficiency	% total plant waste
Decrease consumptions rate	Average material consumption "Kg/Ton produced"
Learning and growth	Number of employees leaving

Example : Discovery of Association Rules

We run association rule Apriori Algorithm to discover the relations between KPIs. Apriori Algorithm uses status as inputs. Below are samples of discovered rules.

- If "Number of new customers" is in the red zone. Then the "Monthly shipping ton compared to same month last year" is in the red zone.
- If "Number of new customers" is in the blue zone. Then the "Monthly shipping ton compared to same month last year" is in the blue zone.
- If "% of orders delivered on-time" is in the blue zone and "% Total plant waste" is in the yellow zone. Then the "Monthly shipping "Ton" compared to same month last year" is in the blue zone.
- If "Average material consumption "Kg/Ton produced" is in the red zone and "Number of New customers" is in the yellow zone. Then the "Monthly shipping "Ton" compared to same month last year" is in the red zone.

From these rules the relations between objectives can be discovered and the strategy map can be redrawn as Fig. 7.4.2 "Increase market share" Objective, "Improve customer satisfaction" Objective and "Increase Machine Efficiency" Objective lead to "Increase profit" objective. Also "Increase Machine Efficiency" and "Decrease Consumptions Rate" are lead to "Improve customer satisfaction" Objective "Sustain Employee Loyalty". Objective supports "Decrease Consumptions Rate" objective.

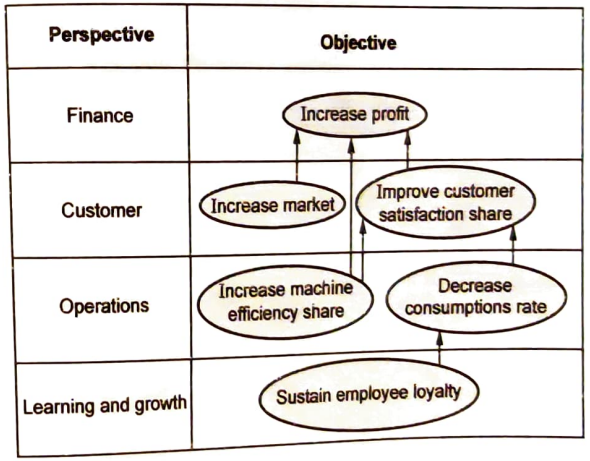


Fig. 7.4.2 Corporate balanced scorecards

7.4.1.1 What are the Good Balanced Scorecards ?

Good balanced scorecards might be said to have good representation on good quality business drivers or KPIs. Qualities of good KPIs include;

- Valid and agreed upon : Drivers must be valid and agreed upon by stakeholders.
- Specific and measurable : Drivers must be specific and measurable systematically.
- Reliable : Information used as KPIs must be reliable.
- Relevant : Drivers must be relevant to business.
- Achievable : Targets assigned for drivers must be achievable.
- Easily understood : Drivers should be easily understood by users timely.

7.13 Fraud Detection

The term fraud here refers to the abuse of a profit organization's system without necessarily leading to direct legal consequences. In a competitive environment, fraud can become a business critical problem if it is very prevalent and if the prevention procedures are not fail-safe. Fraud detection, being part of the overall fraud control, automates and helps reduce the manual parts of a checking process. This area has become one of the most established industry/government data mining applications. It is impossible to be absolutely certain about the ligitimacy of and intention behind an application or transaction. Given the reality, the best cost effective option is to tease out possible evidences of fraud from the available data using mathematical algorithms.

Application of data mining techniques to fraud analysis. Present some classification and prediction data mining techniques which we consider important to handle fraud detection. There exist a number of data mining algorithms and we present statistics-based algorithm, decision tree based algorithm and rule-based algorithm. We present Bayesian classification. Model to detect fraud in automobile insurance. Naive **Bayesian visualisation is selected to analyze and interpret the classifier predictions.**

Instance	Name	Gender	Age-driver	Fault	Driver_rating	Vehicle_age	Output
1	David	M	25	1	0	2	legal
2	BeauJackson	M	32	1	1	5	fraud
3	JeremyDejean	M	40	0	0	7	legal
4	RobertHoward	M	35	1	0.33	1	legal
5	CrystalSmith	F	22	1	0.66	8	legal
6	ChibuikePenson	M	36	0	0.66	6	legal
7	CollioPyle	M	42	1	0.33	3	legal
8	EricPenson	M	39	1	1	2	fraud
9	KristinaGreen	F	29	1	0	4	legal

10	JerrySmith	M	33	1	1	5	legal
11	MaggieFrazier	F	42	1	0.66	3	legal
12	JustinHoward	M	21	1	0	2	fraud
13	MichaelVasconi	M	37	0	0.33	4	legal
14	BryanThompson	M	32	1	0.33	4	legal
15	ChrisWilson	M	28	1	1	6	legal
16	MichaelPullen	M	42	1	0	5	legal
17	AaronDusek	M	48	1	0.33	8	legal
18	BryanSanders	M	49	1	0	3	legal
19	DerekGarrett	M	32	0	0	3	legal
20	JasmineJackson	F	27	0	1	2	legal
X	CrystalSmith	F	31	1	0	2	?

The classifier has to predict the class of instance to be fraud or legal.

P(fraud) = Si/S = 3/20 = 0.15

P(legal) = Si/S = 17/20 = 0.85

Probabilities associated with attributes

Attribute	Value	Count		Probabilities	
		legal	fraud	legal	fraud
Gender	M	13	3	13/17	3/3
	F	4	0	4/17	0/3
age_driver	(20, 25)	3	0	3/18	0
	(25, 30)	4	0	4/18	0
	(30, 35)	3	1	3/18	1/2
	(35, 40)	3	1	3/18	1/2
	(40, 45)	3	0	3/18	0
	(45, 50)	2	0	2/18	0
fault	0	5	0	5/17	0

	1	12	3	12/17
			1	6/17
driver_rating	0	6		3/17
		5	0	5/17
	0.33		0	1/3
		3		3/17
	0.66		2	0
	1	3		3/17
				2/3

7.4.3 Click Stream Mining

In web site are a lot of interconnected web pages that are developed and maintained by a developer or software company. Web mining studies analyzes and reveals useful information from the web. Web mining deals with the data related to the web in website, they may be the, data actually present in web pages. The web can be viewed as the largest unstructured data source available, although the data present on the web sites, basically purpose is it composed them, is structured. This presents a challenging task for effective design of and access to the web pages. Web mining is a term used for applying data mining techniques to web access logs. Data mining is a non trivial process of extracting previously unknown and potentially useful knowledge from large databases. Scientists and engineers want to extract information from it, in order to better understand and to improve its features applied data mining techniques on the web. Therefore, web mining can be defined as the application of data mining techniques to the web related data web mining can be divided into three categories: Web content mining, web structure mining and web usage mining. Web content mining is the process of extracting knowledge from documents and content description. Web structure mining is the process of obtaining knowledge from the organization of the web and the links between web pages. Web usage mining analyzes information about website pages that were visited which are saved in the log files of internet servers to discover the previously unknown and potentially interesting patterns useful in the future. Web usage mining is described as applying data mining techniques on web access logs to optimize web site for users. Click to stream means a sequence of web pages viewed by a user; pages are displayed one by one on a row at a time. Which are saved on the web and proxy servers, when users are visiting them is rapidly becoming one of the most important activities for companies in any sector as most businesses become e-business. Click stream analysis can reveal usage patterns on the company's web site and give a highly imporved understanding of customer behavior. This understanding can then be utilized for improving customer satisfaction with the web site and the company in general, yielding a huge business advantage analysis of clicks is the process of extracting knowledge from web logs This analysis involves first the step of data processing and then applying data mining techniques. Data preprocessing involves data extraction, cleaning and filtration followed by identification of their sessions. Due to the immense

volume of internet usage and web browsing in recent years, log files generated by web servers contain enormous amounts of web usage data that is potentially valuable for understanding the behavior of website visitors. This knowledge can be applied in various ways, such as enhancing the way that the web pages are inteconnected or for increasing the sales of the commercial web sites.

7.4.3 Two Types of Click Stream Data

Transaction Data

Distinguish transactions based on the purpose, the next step could be to perform association and sequence analysis to get insight in visited web sites and customer paths. More sophisticated analysis need more information about transactions so new variables have to be calculated from the log file : number of clicks, average time per site per day, weekday of visit, transaction time in business time, free time, night time, server, type of browser, number of sites visited and which sites were visited in which order, first, second page as well as last, second to last page etc.

Customer Based Data

This information is given in a data warehouse where all customer characteristics and historical information about click behavior. Are stored, to combine this information with the transactional data the users must identify themselves when visiting the web site so the cookie id could be matched with their names and the transactional data can be merged with customer-relevant data.

Benefits Click Stream

Clickstream information can be derived from "raw" click stream data. Here we have tried to discuss some of the most important uses inspired by the needs of the company's active in e-commerce. An important use of clickstream information is personalization of the site to the individual user's needs, delivering services and advertisements based on user interest, and thereby improving the quality of user interaction and leading to higher customer loyalty.

7.4.4 Market Segmentation

Market segmentation is the segmentation of markets into homogenous groups of customers, each of them reacting differently to promotion, communication, pricing and other variables of the marketing mix. Market segments are formed in a way to show the differences between customers within each segment. Thus, every segment can be addressed with an individually targeted marketing mix a market segment consists of a group of customers who share a similar set of needs and wants. It is the marketer's task to identify the segments and decide whom to target. Niche marketing is a more

narrowly defined customer group seeking distinctive mix of benefits. Marketers usually identify niches by dividing a segment into sub segments. The customers in the niche have a distinct set of need for which they are ready to pay a premium to the firm that best satisfied their needs. The niche is not likely to attact other competitors; it gains certain economies through specialization. Niche always has size, profit and growth potentials. Segments are fairly large and normally attract many competitors, whereas niches are fairly small and normally attract only one or two competitors. Grassroots marketing or local marketing activities concentrate on getting as close and personally relevant to individual customers as possible. The ultimate level of segmentation leads to customized marketing or one-to-one marketing. However not every company will go for customization as it leads to raise the cost of product or service by more than the customer is willing to pay. The major segmentation variables are geographic, demographic, psychographic and behavioral. It may be that the top priority of the business is propsecting for new customers or retaining existing customers or cross selling to existing customers. Different business priorities will tend to make some dimensions of segmentation more important than others.

7.4.5 Segmentation Marketing

Customer acquisition

Marketers can use data mining methods to discover the attributes that can predict customers' responses to offers and promotion programs. Thus customer acquisition is possible by matching to those attributes for converting non-customer to respond to new offers and promotions. In segmentation normally the group of people is of similar needs, characteristics or behaviors.

Customer abandonment

Data mining can be used to found out whether a customer has a negative impact on the company's bottom line. It will be better for the company to reject such customer which cost adversely rather than contributing to the growth. In case of segmentation normally targeting is done.

Customer retention

Accordingly the company can design special offers and incentives to retain such customers sufficing there requirements. In market segmentation, as long as the offers are suitable to the specific segment no such problem of customer retention arises.

Market basket analysis

Retailers and direct marketers can use market basket analysis technique of data mining to find product affinities. By identifying the associations between products purchase in point of sale transaction retailers can develope focused promotion strategies.

Segmentation uses positioning technique. Positioning is selecting the marketing mix of products most appropriate for the target segment(s).

7.4.6 Retail Industry

Data mining plays an important role in the retail industry also. Retail industry involves large amount of data that includes transportation, sales and consumptions of goods and services. This data grows rapidly due to increase in purchase and sales into the business. These days, E-commerce is growning fast with the growth of companies and also efficient the online experience. E-commerce describes the buying and selling of products, services, and information via internet. It can be viewed as a business concept that handles eassy and efficiently all previous business management and economic concepts. E-commerce streamlined the business processes, flatter organizational hierarchies and inter-firm collaboration. The databases have become treasure and essential e-commerce tools. To take advantage of the data base, data mining must be integrated into the e-commerce systems. Retail data mining helps in easy to identifying behavior of customer, shopping patterns and distribution policies. As retail data in a very large in quantity, so we design data warehouse to store this large data and effective analysis of data. The main decision has to take while designing the data warehouse is dimension, level and preprocessing to perform the quality and efficient data mining. The main requirement of the retail industry is timely information regarding the customer requirements, trends of market with the cost, profit and quality etc. To get these information effectively and efficiently, powerful multidimensional analysis and visualization tools are required. The retail industry also do some sort of advertisements, give discounts to customers to attact them. An effective analysis is required to get information to check the effect of advertisement on the sales. This analysis can be done by checking the amount of sale done during the sales period before the advertisement and the amount of sale after the advertisement. The sequential pattern mining used to find the change in the customer consumption, adjustment of price and variety in various goods in order to attract customer. Association analysis also helps to get information in order to promote sales. In this way, various data mining tools help in the retail industry.

7.4.7 Telecommunication Industry

Telecommunication can apply data mining for customer retention, fraud analysis, and churn management. It helps to identify telecommunication patterns, catch fraudulent activities, intrusion detection, make better use if resource and improve service quality. Telecommunication industry has been growing very fast as the technology grows. These days telecommunication services has grown from local and long distance voice communication services to fax, pager, cellular phones and e-mails. Now the telecommunication services have integrated with the computer, internet and network and

with other communication technonologies. Due to the advancements in telecommunication technologies and to work these technologies effectively, data mining techniques integrated with these technologies to produce effective results. Data mining helps to identify telecommunications patterns, fraud activities and also helps to better use of resources and improve the quality of services. Data mining improves the telecommunication services in the following ways:

- Telecommunication data involves type of call, location of caller, location of called, time of call and duration of call etc. Multidimensional data analysis helps to identify and compare the system load, data traffic and profit. Analyst can view the charts and graphs of calling resources, destinations etc by using the visualization tools of data mining.
- The main problem faced by the telecommunication industry is due to the fraudulent activities. These fraudulent activities may involve fraudulent calls during busy hour, periodic calls etc. These activities may effect on the performance of the communication network. Data mining methods such as cluster analysis, outlier analysis helps to detect fraudulent patterns and improve the effeciency of the communication services.
- The association and sequential pattern analysis helps to promote various telecommunication services.
- Visualization tools such as association visualization, clustering and association visualization shows very useful telecommunication data analysis. The widespread changes in the adoption and utilization of new technologies in business, even small business has large number of financial transaction. It's the responsibility of the analyzer to analyze these transactions to detect frauds and errors in financial transactions. Due to change in business trends, it's very difficult and complicated to analyze financial transactions by manual methods. Due to limitations of in manual analysis of complex data, we use data mining tools and techniques in various areas to get effective results.

7.4.8 Banking and Finance

Banking industry has hugely benefited from the advancements in digital or digital system. Concept of data stored at branches has given way to centralised databases in different place. Number of channels to access bank accounts has multiplied. Banking systems have become technically strong and customer oriented with online transactions, electronic wire transfers, ATM and cash and cheque deposit machines. As number of channels has increased so is the number of transactions and the related data stored. So currently banks have huge electronic data repositories in their computing storage systems. Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and

market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts. The managers of the banks may go a step further to find the sequences, episodes and periodicity of the transaction behavior of their customers which may help them in actually better segmenting, targeting, acquiring, retaining and maintaining a profitable customer base. Business intelligence and data mining techniques can also help them in identifying various classes of customers and come up with a class based product and / or pricing approach that may garner better revenue management as well.

7.4.9 Risk Management

Today's major challenge in the banking sector is therefore the implementation of risk management systems in order to easy identify, measure, and control business exposure. Here credit and market risk present the biggest challenge, one can observe a major change in the area of how to measure and how to deal with them, based on the advent of advanced database and data mining technology.

Financial Market Risk

Financial instruments stock indices, interest rates, and market risk measurement is based on models depending on a set of underlying risk factor, such as interest rates, stock indices, or economic development. One is interested in a functional form between instrument risk and underlying risk factors as well as in functional dependency of the risk factors itself.

Credit Risk

- Credit scoring/credit rating. Assignment of a customer or a product to risk level.
- Behavior scoring/credit rating migration analysis. Valuation of a customer's or product's probability of a change in risk level within a given time.

Trading

Trading is based on the idea of predicting short term movements in the price or value of a product (currency/equity/interest rate). The goal of this technique is to spot times when markets are expensive or cheap by identifying the reason that are important in determining market returns. Basically in trading system examines the relationship between relevant information and piece of financial assets, and gives you buy or sell recommendations when they suspect an observation.

Portfolio Management

Portfolio level quantifies the risk of a set of instrument or customer including diversification effects. On to the other hand, forecasting models give an induction of the

expected return or price of a financial instrument. Both make it possible to manage firm wide portfolio actively in a risk/return efficient manner.

7.4.10 Customer Relationship Management (CRM)

Customer Relationship Management (CRM) refers to the methodologies and tools that help business manage customer relationships in any organized way. Customer relationship management simply means managing all customer interactions which requires using information about your customers and prospects to more effectively interact with your customers in all stages of your relationship with them. The essence of the information technology revolution and, in particular, the world wide web is the opportunity to build better relationships with customers than has been previously possible in the offline world. Until recently most CRM software has focused on simplifying the organization and management of customer information. Such software, called operational CRM, has focused on creating a customer database that prevents a consistent picture of customer's relationship with the company, and providing the information in specific applications. However, the sheer volumes of customer information and increasingly complex interactions with customers have propelled data mining to the forefront of making your customer relationshp profitable. Data mining is the process that uses a variety of data analysis and modeling techniques to discover patterns and relationships in data that may be used to make accurate predictions. It can help you to select the right prospects on whom to focus, offers the right additional products to your existing customers and identify good customers who may be about to leave. CRM applications that use data mining are called analysitc CRM. By combining the abilities to respond directly to customer requests and to provide the customer with a highly interactive, customized experience, companies has a greater ability today to establish, nurture, and sustain long term customer relationships than ever before. The ultimate goal is to transform these relationships into greater profitability by increasing repeat purchase rates and reducing customer acquisition costs. The needs to better understand customer behavior and focus on those customers who can deliver long-term profits have changed how marketers view the world. Basic model of CRM contains a set of following components.

i) Creating a customer database

- **Transactions** : This should include a complete purchase history.
- **Customer contacts** : There are an interesting number of customer contact points from multiple channels and contexts.
- **Descriptive information** : This is for segmentation and other data analysis purposes.

- **Responses to marketing stimuli** : This part of the information file should contain whether or not the customer responded to a direct marketing initiative, a sales contact, or any other direct contact.

ii) Analyzing the data

Basically its different type of statistical methods ranging such as cluster and discriminate analysis have been used to group to gather customers with similar behavioral patterns and descriptive data which are then used to develope different product offerings or direct marketing campaigns.

iii) Customer selection

After the customer database has been created and then analyzed, the next step is to consider which customers to be target? The results from the analysis could be of various types.

iv) Targeting the customers

More conventional approaches for targeting selected customers include a portfolio of direct marketing methods such as telemarketing, direct mail, and, when the nature of the product is suitable, direct sales, television, radio, or print advertising are useful for generating awareness.

v) Relationship programs

It is more technique for implementing CRM, that customers match realizations and expectations of product performance, and that it is critical for them to deliver such performance at higher and highest levels as expectations increase due to intense or heightened competition, and changing customer needs.

7.5 Data Analytics Lifecycle

Analytics process best practices sequence of discovery to project completion draws from actual methods in the reduce of data analytics and decision science. This synthesis was developed after collection input from data scientists and consulting established approaches that provided input on pieces of the process. Several of the processes that were consulted include these different phase in data analytics lifecycle.

Phase 1 Discovery

In this phase now project development team assesses the resources available to support the project in terms of people, technology, time and data.

Phase 2 Data preparation

After collection data the team needs to execute extract, load, and transform or extract, transform and load (ETL) to get data into the sandbox.

Phase 3 Model planning

In this phase project development team determines the methods, techniques and workflow it intends.

Phase 4 Model building

Project team develops data sets for testing, training and production purpose.

Phase 5 Communicative results

Main and important phase collaboration with major stakeholders, determines if the results of the project are a success and frailer with their reason.

Phase 6 Operationalize

Project to implement the models in a production environment

7.6 Introduction to Big Data Business Analytics

Today's here are different type of many big data analytical tools available, such as predictive analytics, descriptive analytics, and survival analysis. More than methods and techniques are designed, such as linear regression, support vector machines, logistic regression, and neural networks. They are directly applied in many business areas.

Retail

The changes of the marketplace and the diverse demands of consumers lead to the increasing complexity of data volume and different data type. One main purpose is that the personal opinions contained in the raw data collected for analysis purpose are unstructured data frequently. Other contributors including online shopping, social media and collaboration also intensify the burden of data analyzing.

Marketing

Customer classification is helpful for designing compaigns and customizing promotion.

Fraud detection

Fraud detection is used in many areas including credit card fraud is one type of fraud, insurance claim fraud, and money laundering and tax evasion fraud. Basically, it is to detect anomalies from data and transactions. Supervised, unsupervised, and social network learning can be used for fraud detection.

Customer relationship management

The cost of retaining customers is significantly lower than the cost of replacing them, making the ability to identify customers at risk of churning vital. Key performance

indicators are used to describe customers, including demographic information and recent call patterns for each individual customer. Predictive models based on these fields use changes in customer call patterns that are consistent with call patterns of customers who have churned in the past to identify people having an increased churn risk.

Social network analysis

The increasing use of social networks, such as Facebook, Twitter, and Weibo (http://www.facebook.com) has produced and is producing huge volume of data. For example: Twitter posts more than 1000 million tweets every day. Weibo is reported to have over 500 million active users per day. Business firms and other organisations are interested in discovering new this type of social network business insight to increase business performance. By using it advanced analytics, enterprises can analyze big data to learn about relationships underlying social networks that characterize the social behavior of individuals and groups or more than groups. Using data describing the relationships we are able to identify social leaders who influence the behaviour of others in the network, and on the other hand, to deterine which people are most affected by other network to bee participants. We can also use diffusion analysis to identify the individuals most affected by the group leaders and target the marketing to them.

7.7 State of the Practice in Analytics Role of Data Scientists

Analytics role of data science solely deals with getting insights from the data whereas analytics also deals with about what one needs to do to 'bridge the gap to the business' and 'understand the business priorities'. Provides subject matter expertise for analytical techniques, data modeling, and applying valid analytical techniques to given business problems. Ensures overall analytics objectives are met. Designs and executes analytical methods and approaches with the data available to the related project. It is the study of the methods of analyzing data to data, ways of storing it, and ways of presenting it. Often it is used to describe cross field studies of managing, storing and analyzing data combining computer science, statistics, data storage, and cognition. It is a new field so there is not a consensus of exactly what is contained within it.

(See Fig. 7.7.1 on next page.)

Data science is a combination of mathematics, programming, statistics, the context of the problem being solved easy, ingenious ways of capturing data that may not be being captured right now plus the ability to look at things 'differently' and of course the significant and necessary activity of cleansing, preparing and aligning the data.

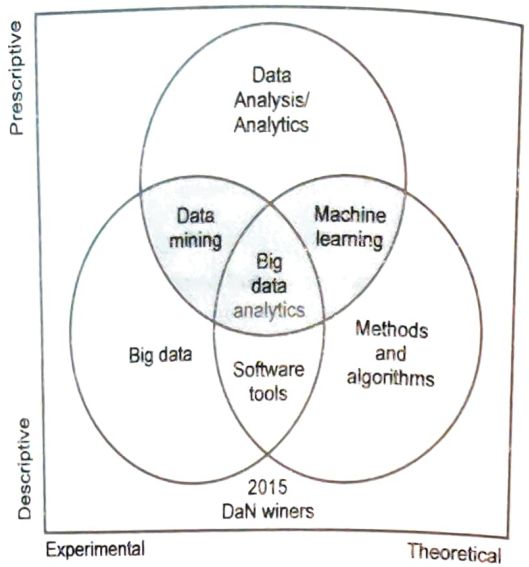


Fig. 7.7.1 State of the practice in analytics role data scientists

7.8 Key Roles for Successful Analytic Projects

Provides subject matter expertise for analytical techniques, data modeling and applying valid analytical techniques to given business problems.

Business user

Usually a business analyst, line manager, or deep subject matter expert in the project domain fulfills this role.

Project sponsor

Generally provides the funding and gauges the degree of value from the final outputs of the working team.

Project manager

Ensures that key milestones and objectives are met on time and at the expected quality of projects.

Database administrator

Provisions and configures the database environment to support the analytics needs of the working team.

Data engineer

Technical skills to assist with tuning queries for data management.

7.9 Main Phases of Life Cycle

Main phase of big data life cycle or main phases of life cycle model it can be used in an iterative manner with a project entering the life cycle at any point. Datasets in the same project may be situated at different phases of the cycle.

Describe activity

Describing the data and process used in the analysis every and each step capturing the provenance trace is crucial for big data. The earlier curation-related tasks are being planned in the data management life cycle, the easier they may be to execute take for instance the collection of metadata for datasets.

Assure activity

Describing the data, including its processing and then analysis, from the start of the projects assure play mail role toward assuring the quality of the whole data. It's not in itself sufficient. Basically assuring the quality of data includes quality assurance, which may occur once a dataset is analyzed and quality control, a pro-active process with procedures in place to ensure quality. Data quality is directly related to the veracity characteristic of big data, as accuracy, completeness, and uncertainty about sources play a crucial role in veracity. Similar to data documentation, issues of quality arise at every stage of the life cycle, planning, including acquisition, preparation, analysis and preservation and discover.

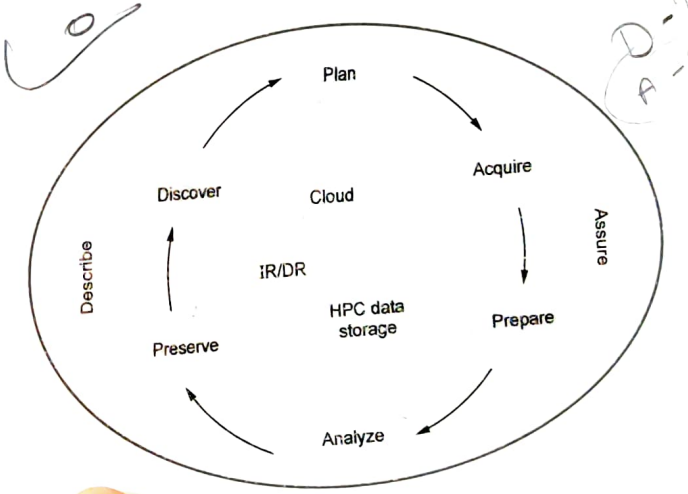


Fig. 7.9.1 Main phase of life cycle

Planning activity : Preservation must be discussed at the planning stage due to the potential volume of data to be preserved.

Acquire activity : In acquire activity reflects how data is produced, generated, and ingested in the research process

Prepare activity : Preparing datasets and staging them for analysis is a time consuming step with big data and its complexity is often overlooked.

Analysis activity : The analysis activity is the domain of the scientists performing research

Preserve activity : In order to preserve results for long-term use.

Discover activity : In discover activity refers to the set of procedures that ensures that datasets relevant to a particular analysis can be found by other than those involved in the project.

7.10 Developing Core Deliverables for Stakeholders

Stakeholders are the originator of the project management organization that is responsible for the delivery of stakeholders' expectation and satisfaction. The successful delivery of any project deliverables highly depend on stakeholder engagement and management and the effective engagement and management of stakeholder relies on project manager's ability to identify stakeholders' expectations from the beginning to close-up. Researchers described project stakeholders management as a process in which project team facilitates the needs of stakeholders to identify, discuss, agree and contribute to achieve their objectives, describes stakeholder relationship management through six continues processes, including identifying stakeholders, analyzing, engaging, identifying information flow, enforcing stakeholder agreement, and stakeholder debriefing. Additionally, form the base-organization viewpoint suggested three processes of stakeholder identification, assessement and prioritisation. Developing core deliverables has drawn the key stakeholder management processes from the literature to construct its mediating factor. The mediating variable of manage-through-stakeholder consists of five main observed variables of stakeholder identification and classification, communication, engagement, empowerment, and risk control. The aim here is to investigate the mediating role of manage-through-stakeholder on the relationship between stakeholder influential variables and project success. Therefore, and in line with the research objectives, and the hypothesized model, we have treated manage-through-stakeholder construct as a latent factor with five observable indicator is as followed.

H2a : The relationship between stakeholders' power and project success is significantly improved when manage-through-stakeholder is mediated.

H2b : The relationship between stakeholders' interest and project success is significantly improved when manage-through stakeholder is mediated.

H2c : The relationship between stakeholders' legitimacy and project success is significantly improved when manage-through stakeholder is mediated.

H2d : The relationship between stakeholders' urgency and project success is significantly improved when manage-through-stakeholder is mediated.

H2e : The relationship between stakeholders' proximity and project success is significantly improved when manage-through-stakeholder is mediated.

H2f : The relationship between stakeholders' relationship network and project success is significantly improved when manage-through-stakeholder is mediated.

7.11 Reason for use Data Mining for Business Intelligent Application

1) Supporting decision making network and query analysis
2) The capability of organization
3) Comparision other organization as per product in market
4) Easy to manage.

7.11.1 Advantage of Business Intelligence

i) Reduced employ costs
ii) Make detailed data for actionable purpose
iii) Maintain customer relationship profile
iv) Decision making
v) Timely and faster decision
vi) Analysis of customer validation need
vii) Reduced human intervention resulting in fewer human errors.
viii) Reduced cycle times
ix) Maintain better quality data
x) Incomplete business process are immediately find before downstream process can use corrupt data.
xi) Widely or broad applicable handles many attribute, large amount of data.
xii) The capabilities of the organization.

7.12 Business Intelligence Include

Business intelligence include tools in different different categories are as following

i) Business planning purpose
ii) Re-engineering
iii) Web mining
iv) Text mining
v) Human resources
vi) Associative query logic
vii) Executive Information System (EIS)
viii) Management Information System (MIS)
ix) Decision Support System (DSS)
x) Document warehouse
xi) Real time business intelligence
xii) Dynamic or chain demand management
xiii) Dashboarding and information visualization
xiv) Customer Relationship Management (CRM)
xv) Improve marketing
xvi) Datamining
- Intelligent analysis
- Finance and budgeting
- Online Analytical Processing (OLAP)

7.12.1 Advance Topics of Data Mining and its Application

7.12.1.1 Mining-Time Series and Sequence Data

Time series database contain time related data such stock - market data or logged activities these database usually have a continuous flow of new data coming in which sometime cause its need for a challenging real time analyse, data mining in such database commonly includes the study trends and correlation between evolutions of different variable, as well as the prediction of trends and movement of the variable in time.

Different - different type of methodologies are as follows
i) Random sampling
ii) Sliding windows

iii) Histograms
iv) Multi resolution methods
v) Sketches synopses

i) Random sampling :

- Random sampling is the purest form of sampling.
- It is a statistical sampling method is simple random sampling.
- This sampling can, each item in the population has the same probability to being selected as part of the sample as any other item base.
- Where in number of total set of observations that can be made is called the population.
- In sampling method is basically procedure for selecting sample element from a population.
- Simple random is simple are self-weight based.

Sampling is two type :

a) Simple random sample with replacement
b) Simple random sample without replacement.

ii) Sliding window :

1) Sliding window model to be analyze stream data.
2) Sliding window is attractive because of its great simplicity intuitiveness and particularly the fact that it is an online algorithm.

iii) Histogram :

1) Histogram can be use binning to approximate data distribution and are a popular from of data reduction purpose.
2) When Histogram an attribute B, partitions the data distribution of B into disjoint subset, if each partition subset can represent only a single attribute value, this single attribute are called singleton.

There are several attribute value partitioning rules some of them are as follows

1) Equal width : In an equal width histogram.
2) Equal frequency : Is an equal frequency histogram.
3) V - Optimal : Is the one with the least variance.
4) MaxDiff : Consider the difference between each pair of adjacent value.

4) Multi - resolution methods :

- Multi - resolution method to be data reduction methods, it can be reduce large amount of data.
- Data reduction method is the use of divide - and - conquer strategies such as multiresolution data structure.

5) Sketches :

- Sampling techniques and sliding window model both are handle only small data.
- It can be handle multiple level of detail.
- Multiple pass over the data, such as handle histograms, wavelets and sketches methods.

7.12.2 Mining Text Database

7.12.2.1 Introduction

- A text is any sequence of symbol to be drawn from an alphabet. In a large portion of any information available worldwide in electronic form is actually in text from.
  Example : Natural language text, book, journal, newspaper, jurisprudence database.
- Basically text database is an online platform to be original text and translations. This original text can be upload into be text database.

7.12.2.2 Information Retrieval

Information retrieval and database system can handle to be different different type of data, there are some database system problem that are usually not present in information retrieval system. Like to concurrency control, recovery transaction management and update unstructed document, approximate search based on key work and the motion of relevance.

Measures in Text Retrieval

There are two type of measures are as follows :

1) Precision : Retrieval document that are in fact relevant to the query

$$\text{Precision} = \frac{|\{\text{Relevant}\} \cap \{\text{Retried}\}|}{|\{\text{Relevant}\}|}$$

2) Recall : Retrieval document that are in fact relevant to the query were in fact.

$$\text{Recall} = \frac{\{\text{Relevant}\} \cap \{\text{Retrivel}\}}{|\{\text{Retrived}\}|}$$

Syntactic Search

Syntactic searching : When desired outcome of a search pattern is clear so the main concern of text database is efficiency. This are two type of possible to pattern matching.

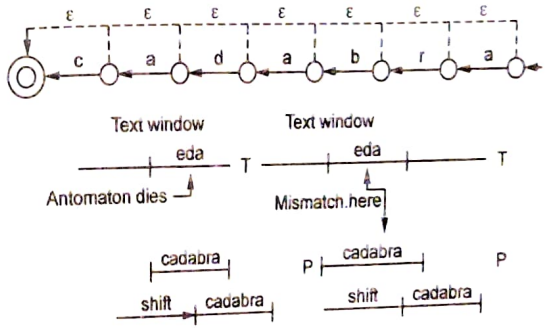
i) Sequential searching : Assumes that it is not possible to preprocess the text, so only the pattern is preprossed and then the whole text database is sequentially scanned.

ii) Indexed searching : Preprocess the text so as to build a data structure over it an index. Which can be used later to speed up searches.

7.12.3 Keyword Based Retrieval

Keyword - based retrieval, a document is represented by a string pattern P of m character, a text T of n character and required to point out all the occurrences of P in T.

For example : A NFA based example



7.12.4 World-Wide-Web

The world-wide-web (WWW) and its associated information services system it can be consist of millons of clients and servers for accessing linked text document. WWW became publicly available on the Internet on 1991. Such as value. Google, America, Online and Vista provide rich world wide online information serives where data objects are liked togather to facilitate interactive access user seeking information of interest travers from one object links to another such system provide sample opportunities and challenges for data mining.

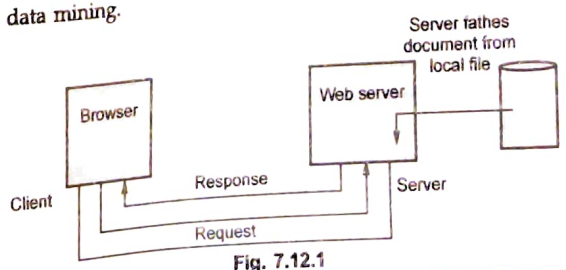


Fig. 7.12.1

i) Browser : It is software, which responsible to properly displaying a document there are two type of browsers.

i) Graphical ii) Text

ii) Web-server : Web-server is designed to serve HyperText Transfer Protocol (HTTP)

Example : i) www.yahoo.com ii) www.google.com

7.12.4.1 Characteristics of Web-Application Development

1) Application Related.

i) Content ii) Hyper

2) Development Related

i) Legacy integration ii) Process iii) Development team

3) Usage Related

i) Natural context ii) Unpredictable iii) Diversity iv) Magnitude of use base

7.12.5 Business Intelligence Current use in Data Mining

i) Clustering : Cluster topology over a period of time will provide valuable information for the better useful of the usage pattern.

ii) Association : Identify relationships

iii) Classification : Determine or find a set of function that describes and distinguish data class.

iv) Text mining : The process of extracting document information.

v) Segmentation : Task process divide in small groups.

vi) Rough set approach : Reduce to large process based on similar case.

vii) OLAP : Mining knowledge in multidimensional.

Review Questions

1. *Explain application of business intelligence.*
2. *What is business intelligence in datamining and its application ?*
3. *Explain advantage of business intelligence.*
4. *Explain data analytic lifecycle.*
5. *What is the advantage of busi... ...ence.*

□□□

8 Advance Topics

Syllabus

Introduction and basic concepts of following topics. Clustering, Spatial mining, Web mining, Text mining,
Big data : *Introduction to big data : Distributed file system - Big data and its importance, Four Vs, Drivers for big data, Big data analytics, Big data applications. Algorithms using map reduce, Matrix-Vector multiplication by map reduce. Introduction to Hadoop architecture : Hadoop architecture, Hadoop storage : HDFS, Common hadoop shell commands, Anatomy of file write and read., NameNode, Secondary NameNode and DataNode, Hadoop mapreduce paradigm, Map and reduce tasks, Job, Task trackers - Cluster.*
Setup - SSH and Hadoop configuration - HDFS administering - Monitoring and maintenance.

Contents

8.1 Clustering

Clustering is the process of grouping a collection of objects (usually represented as points in a multidimensional space) into classes of similar objects. Cluster analysis is a very important tool in data analysis. It is a set of methodologies for automatic classification of a collection of patterns into clusters based on similarity. Intuitively, patterns within the same cluster are more similar to each other than patterns belonging to a different cluster. It is important to understand the difference between clustering (unsupervised classification) and supervised classification. Cluster analysis has wide applications in data mining, information retrieval, biology, medicine, marketing and image segmentation. With the help of clustering algorithms, a user is able to understand natural clusters or structures underlying a data set. For example, clustering can help marketers discover distinct groups and characterize customer groups based on purchasing patterns in business. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations. Typical pattern clustering activity involves the following steps :

- Pattern representation (including feature extraction and/ or selection),
- Definition of a pattern proximity measure appropriate to the data domain,
- Clustering,
- Data abstraction, and
- Assessment of output

Requirements of Clustering in Data mining :

- **Scalability** - We need highly scalable clustering algorithms to deal with large databases. Ability to deal with different kind of attributies-Algorithms should be capable to be applied or any kind of data such as interval based (numerical) data, categorical, binary data.
- **Discovery of clusters with attribute shape** - The clustering algorithm should be capable of detect cluster of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small size.
- **High dimensionality** - The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
- Ability to deal with noisy data - Databases contains noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- **Interpretability** - The clustering results should be interpretable, comprehensible and usable.

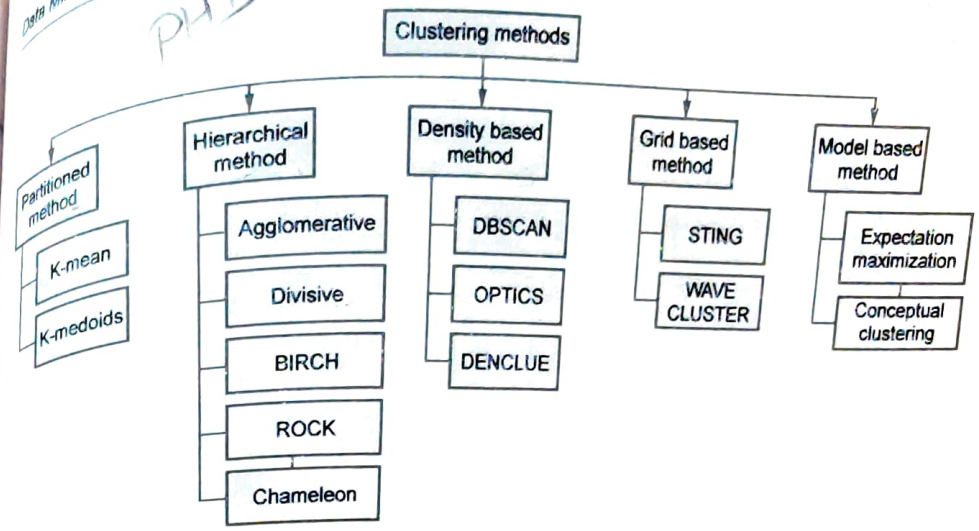


Fig. 8.1.1 Clustering method

Partitioning Method

In this clustering method clustering creates the clusters in one step instead of creating several steps. Only one set of clusters is formed at the end of clustering, although several sets of clusters may be created internally. As we know that only one set of clusters will be formed then n user must have to specify the input (the desired number of clusters.) The most well-known and commonly used partitioning methods are k-means, k-medoids.

- **K-means method :** centroid based method the k-means method takes the input parameter, k, and partitions a set of n objects into k clusters so that the Resulting intra cluster similarity is high but the intercluster similarity is low.
- **k-medoid** method Rather than a reference point or mean value of the cluster, we choose actual objects to represent the clusters, one object per cluster. Each leftover Object is clustered with the choose n object to which it is most similar. Then per formed the partitioning method based on the principal of minimizing the sum of dissimilarities between each object and its corresponding reference point or mean value.

Hierarchical Method

In this Method chain data objects into a treelike structure of cluster. Tree of clusters is called dendrograms. Every cluster node contains child clusters, sibling clusters and separation of the points hooded by their common parent. In general there are two types of hierarchical method

- **Agglomerative Method** : It is a bottom up-approach, each object has their own cluster and these clusters are merged to form a large clusters i.e. a single cluster until some termination conditions are satisfied.
- **Divisive Method** : It is top-down method in which clusters are subdivided into smaller and smaller parts until all part or object creates their own cluster or until they satisfies certain or specific termination condition like a desired number of clusters to be obtained or the diameter of each cluster reach the threshold.
- **BRICH Method** : (balanced iterative reducing and clustering using hierarchies BRICH) is designed for clustering a large amount of numerical data. The basic idea is that a tree is formed that captures needed information. Clustering is performing on the tree itself; the nodes in the tree contains the information which is used for the calculation of distance values. BRICH contains two new concept called Clustering feature (CF) and clustering tree (CE). Both of the CF and CE summarize the cluster representations, and provide helps in achieving good speed and scalability for large database. The CF is three-dimensional vector which contain or summarize information of objects of a clusters which are sufficient to calculate the measurements which help in clustering decisions. Whereas CF-Tree is height balanced tree which store the CF for making hierarchical clustering. It contains Two parameter one is branching factor (BF) which describes the maximum number of child per non leaf node, and other one is threshold(T) which describes the diameter of sub cluster which are stored on the trees leaf node BRICH tries to produce the best possible cluster among all the cluster from the given resources.
- **Chameleon** : It is a hierarchical clustering method which uses the dynamic modeling approach to find out the similarity between the pairs of clusters.

Density Based Method

In this method we find out the arbitrary shapes clusters in which the objects are stored into the data space according to their low density here, the object is called or represented as dense region all the cluster are made on the basis of their low density between these regions.

- **DBSCAN** : density based clustering method based on connected regions It is density based clustering method for handling spatial data with noise in application or data base. It uses the high density region for making the cluster and the other regions which have low density are kept outside the cluster by marking as outlier. There is no need to define the number of clusters in advanced. By using the "Minpt" parameter it is able to find out the cluster which is totally different. "Density reach ability" and "Density connect ability" are the two concepts which are used during making the cluster which in turn have asymmetric and symmetric

relation. "Minpt" and "e" are the two parameters, if point k contains more "Minp" t than the e-neighborhood then a new cluster with core object will be created, then the DBSCAN will gather the density reachable object from these core objects.

- **OPTICS** : Ordering point to identify the clustering structure Optics creates the liner ordering of objects in the database. Like the DBSCAN it use two parameter "e" and "Minpt" where define the maximum distance and "Minpt" define the number of points or objects required to make a cluster. For making clustering automatic and iterative augment ordering of objects in the database is created.
- **DENCLUE** : (Clustering based on density distribution functions) DENCLUE use the density distribution function for making the clusters. It uses the influence function which wedges the data point along with its neighborhood points.The points are arranged in the hill climbing manner where the points having the same local maximum are placed together into the cluster DENCLUE have strong mathematical foundation and good properties which perform the arbitrarily shaped cluster in high dimensional data set with large amount of noise.

Grid Based Method

Grid based method is different from other clustering algorithms. It uses the multi resolution grid data framework. It provides helps in reducing the computational complexity for very large data sets. First it separates the data into finite number of cells then it calculates the density for each cell. Based on the density of each cell sorting is done and center of the clusters are marked neighbor cell are traversed.

- **STINGS** : statistical information grid STINGS rupture the whole spatial area into rectangular cells. These rectangular cells elevate tree like structure which reciprocate to other different level of resolution. Every cell is rupture into other cells at a high level to make the next lower level. This algorithm assumes that a query can be answered from the stored statistical information which is reciprocated in the tree. The upper part of the tree consists the entire space and the lower area or level has one leaf for each smallest cells. In this algorithm only vertical and horizontal boundaries are built. Scanning is done one time and all the parameters like, mean, variance, distribution are determined for each cell which makes it more efficient.
- **WAVECLUSTER** : clustering using wavelet Transformation in this approach every grid cell encapsulate the information of points that is mapped into the cell. This pruned knowledge/information is then applied into the multi resolution wavelet transform for the cluster analysis. This multi resolution property help in recognizing the varying level of accuracy

Model Based Method

In this method observation are done to find out the features of the objects and these feature are engender via the distribution, which have free normal density distribution.

- **Expectation-maximization** : EM is the most preferred iterative refinement method that is used to figure out the parameter estimates. Each cluster is defined by parametric probability distribution. Objects are assigned to cluster according to their mean value with some weight associated with objects.
- **Conceptual method** : In this I method is an unsupervised machine learning method for the Classification of unknown classification. Concept based structure is used to separate the generated classes from the ordinary data.

8.2 Spatial Data Mining

Spatial data mining is the application of data mining techniques to spatial data. Data mining in general is the search for hidden patterns that may exist in large databases. Spatial data mining is the discovery of interesting the relationship and characteristics that may exist implicitly in spatial databases. Because of the huge amounts (usually, tera-bytes) of spatial data that may be obtained from satellite images, medical equipments, video cameras, etc. It is costly and often unrealistic for users to examine spatial data in detail. Spatial data mining aims to automate such a knowledge discovery process. Thus it plays on important role in.

- Extracting interesting spatial patterns and features
- Capturing intrinsic relationships between spatial and non spatial data
- Presenting data regularity concisely and at higher conceptual levels and
- Helping to reorganize spatial databases to accommodate data semantics, as well as to achieve better performance. Spatial data base stores a large amount of space related data, such as maps, preprocessed remote sensing or medical imaging data and VLSI chip layout data. Spatial databases.

Have many features distinguishing them from relational databases. They carry topological and/or distance information, usually organized by sophisticated, multi dimensional spatial indexing structures that are accessed by spatial data access methods and often require spatial reasoning, geometric computation, and spatial knowledge representation techniques.

8.3 Spatial Data Mining Structure

The spatial data mining can be used to understand spatial data, discover the relation between space and the non space data, set up the spatial knowledge base, excel the query, recognize spatial database and obtain concise total characteristic etc.

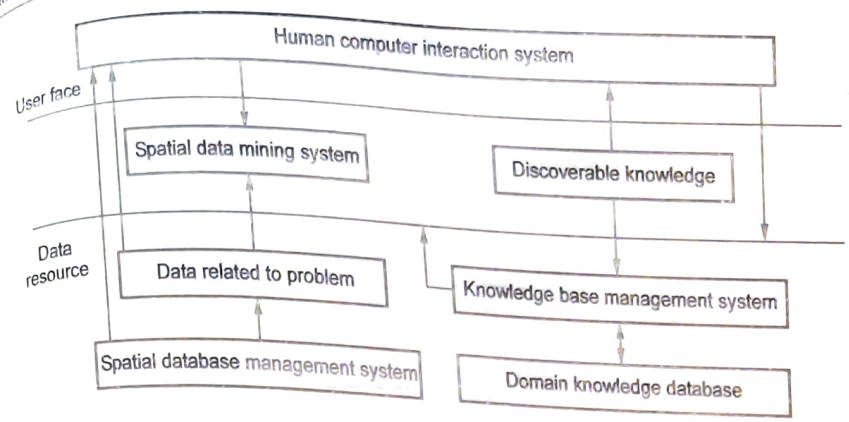


Fig. 8.3.1

The systematic structure of the spatial data mining can be divided into three layer structures mainly, as shown in Fig. 8.3.1. The first layer i.e. customer interface layer is used for input and output. This layer is also known as DB interface or user face. Data is fetched from the storage using the *DB interface* which enables optimization of the queries. The second layer miner layer is mainly used for management of data, selection of algorithm and to store the mined knowledge. For example, it may decide that only some attributes are relevant to the knowledge discovery task, or it may extract objects whose usage promises good results. Third and last data resource layer, which includes the spatial database and other related data and knowledge bases which is original data of the spatial data mining. Data inputs of spatial data mining are more complex than the inputs of classical data mining because they include extended objects such as lines,

Primitives of Spatial Data Mining

Rules : There are several kinds of rules can be discovered from databases in general. For example characteristics rules, discriminate rules, association rules, or deviation and evolution rules can be mined. A Spatial characteristics rules a general description of the spatial data.

Thematic Maps : Thematic map is map primarily design to show a theme, a single spatial distribution or a pattern.

Using a specific map type. these maps show the distribution of features over limited geography areas. Each map defines a partitioning of the area into a set of closed and disjoint regions; each includes all the points with the same feature value. Thematic maps present the spatial distribution of a single or a few attributes. this differs from general or reference maps where the main objective is to present the position of the object in relation to other spatial objects. Thematic maps may be used for discovering different rules.

8.4 Web Mining

The web mining is a combination of the two Singular areas of in progress research initial one is the data mining and Second one is world wide web (WWW). It can be able to be mostly Defined as the finding and investigation of useful information from WWW. Web mining is the make use of data mining performance to without human intervention discover and mine information from Web documents and services. The Web mining research is a come together research area from several research community, such as database, IR, and AI research community especially from machine learning and NLP. This paper is a challenge to put the research done in a more controlled way from the machine learning end of view. However, the methods of the Discover do no repeatedly use well-known machine realize algorithms. Since this is a huge, interdisciplinary, and very active examine area, there are lacking doubt some omission in this treatment Web mining should contain the sub sequent's their sub tasks.

- **Source finding :** The task of mine related information.
- **Information collection and preprocessing :** Selecting and preprocessing necessary Information from related documents.
- **Generalization :** Automatically extract general patterns at single Web sites as well as across multiple sites.
- **Analysis :** Validation and/or interpretation of the mined patterns.

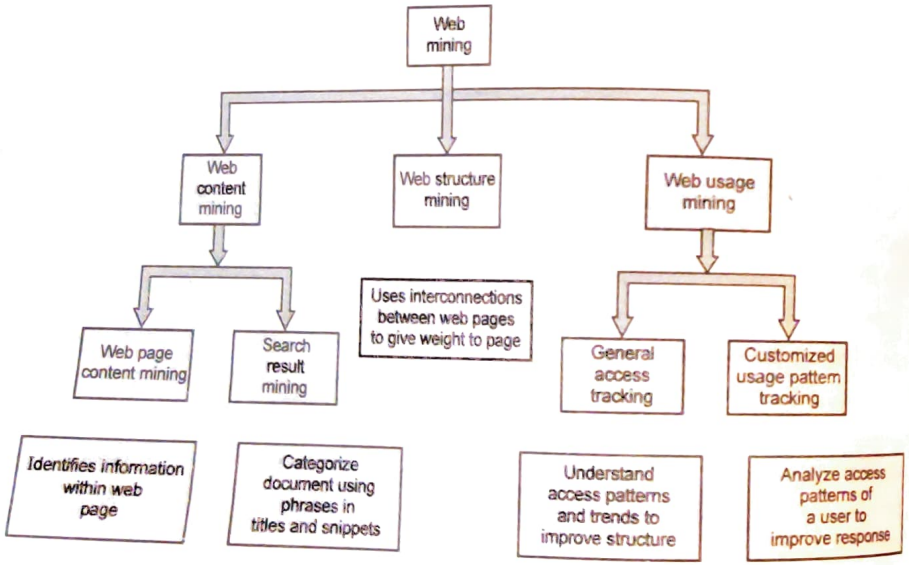


Fig. 8.4.1 Web mining

Web Mining Categories

There are three areas of Web mining according to the usage of the Web data used as input in the data mining process, namely.

Web Content Mining (WCM)

Web content mining performed by extracting useful information from the content of a web page/site. It includes extraction of structured data/information from web pages, identification, match, and integration of semantically similar data. Unstructured text mining approach and Semi-Structured and Structured mining approach.

Unstructured Text Data Mining

Web content data is much of free text data. The research around applying data mining technique to understand text is termed Knowledge Discovery in Texts (KDT), or text data mining, or text mining. Hence, one could consider text mining as an instance of Web content mining.

Semi-Structured and Structured Data Mining

Structured data on the Web are often very important as they represent their host pages, due to this reason it is important and popular. Structured data is also easier to extract compared to unstructured texts. Semi-structured data is a point of convergence for the Web and database.

Communities : The former deals with documents, the letter with data. The form of that data is evolving from rigidly structured relational tables with numbers and strings to enable the natural representation of complex real world objects like books, papers, movies, etc., without sending the application writer into contortions.

Web structure

- Web structure mining tries to discover the model underlying the link structures of the Web. The model is based on the topology of the hyperlink with or without the description of the links. This model can be used to categorize Web pages and is useful to generate information such as the similarity and relationship between different Web sites. Web structure mining could be used to discover authority sites for the subjects (authorities) and overview sites for the subjects that point to many authorities (hubs).

Link-based Classification : The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page.

Link-based Cluster Analysis : The goal in cluster analysis is to find naturally occurring subclasses.

Link Type : There are a wide range of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link.

Link Strength : Links could be associated with weights.

Link Cardinality : The main task here is to predict the number of links between objects. There is many ways to use the link structure of the Web to create notions of authority.

Web usage mining

Focuses on techniques that could predict user behavior while the user interacts with the Web. As mentioned before, the mined data in this category are the secondary data on the Web as the result of interactions. These data could range very widely but generally we could classify them into the usage data that reside in the Web clients, proxy servers and servers. The Web usage mining process could be classified into two commonly used approaches.

Data Collection : During this stage, data are collected either from Web servers or from clients that visit a Web site.

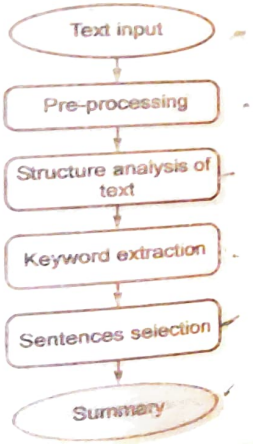
Data Preprocessing : This is the stage that involves primarily data cleaning, user identification and user session identification.

Pattern Discovery : During this stage, knowledge is extracted by applying Machine Learning techniques, such as clustering.

Knowledge Post-processing : In this last stage, the extracted knowledge is evaluated and presented in a form that is understandable to humans.

8.5 Text Mining

As there is huge growth in web, digital libraries, Medical data hence online textual documents are becoming more effective. So extracting knowledge out of these text documents is one of major research area. An effective reader will always generate relationships between texts so that a hypothesis can made. The extraction of useful patterns out of the textual resources is known as Text Mining.



Technology Premise of Text Mining

There are several technology premises for mining the text. Some of them are listed below and later on to describe one by one.

- Summarization
- Information extraction
- Categorization
- Visualization
- Clustering
- Topic tracking
- Question answering
- Sentiment analysis

Summarization : Summarization is one of the most important techniques. In simple words summarization is a process of making summary of any document containing large amount of information while theme or main idea of Document is maintained.

Information Extraction : Information Extraction utilizes relations within the text. It uses pattern matching for it. For example, in RDBMS stored data is available in the form of tables. When data is in unstructured form, information cannot be extracted with ease (no fixed reference). In IE natural language document is converted into structured one and then the knowledge is extracted.

- Entity Extraction (associating nouns with entities)
- Concept Extraction (noun and phrases)
- Token Extraction (characters without separator)
- Term Extraction (token with specific semantic purpose)
- Atomic Fact Extraction (subject with actions)
- Complex Fact Extraction

Categorization : Categorization is a supervised learning technique which places the document according to content. Document categorization is largely used in libraries. Document classification or text categorization has several applications such as cell center routing, automatic metadata extraction, word sense disambiguation

Visualization : Visualization which dates back to two century is a computer graphic effect to represent information and revealing relationships

Clustering : Text clustering is a document's textual similarity based unsupervised technique (does not require training data) which is used by data analysts to divide the text into mutually exclusive groups. Text mining has very less application of clustering. Clustering can be divided into 2 parts- hierarchical clustering and partition clustering.

Topic Tracking : As the name suggests topic tracking is a process of following any specific topic based on requirement or interest. It predicts possible topic of interest for use on the basis of his topic interest history.

Question Answering : Natural language queries or questions answering is responsible to decide a way find a more suitable answer for particular question.

Sentiment Analysis : Sentiment Analysis which is also known as opinion mining is configured of user's emotion, mostly into several classes which are positive, negative, neutral and mixed and mixed. It is mainly used to get people's view or attitude towards anything which includes services and products.

Tools of Text Mining

There are some tools on the internet which can be used for mining the text. These tools follow step by step process for the purpose of mining.

- Import event data into mining tool.
- Linguistic processing.
- Factor analysis.
- Cluster analysis.

8.6 Introduction to Big Data

Different definitions of big data are as following :

Big data is relatively a new concept a several definitions have been given to it by researchers organization and individuals. As far back as the information technology research company Gartner defines : Big data are high- volume, high - velocity and high-variety information assets that require new form of processing to enhance decision making, insight discovery and processing optimizing.

Or

Big Data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high velocity capture, discovery and analysis.

Or

Big data is defined as the representation if the progress, usually include data set with sizes beyond the ability of current technology, method and theory to capture, manage and process the data within a tolerable elapsed time.

Or

Big Data is the collection of massive amounts of information, whether unstructured or structured. Today, many organizations are collecting, storing and massive amounts of

data. This data is commonly to as 'big data" because of its volume, the with which it arrives and the variety of forms it takes.

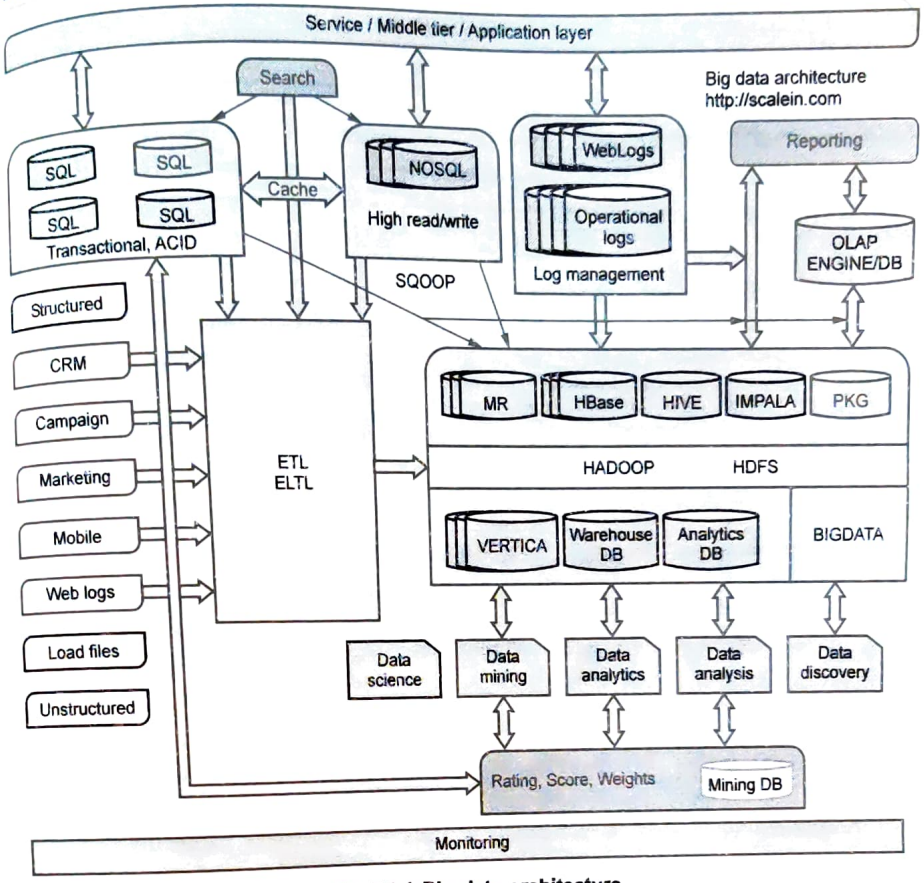


Fig. 8.6.1 Big data architecture

A. Data Source :

True source of data are coming heterogeneous data sources. This is typically your data stores (SQL or NoSQL) that give a structured data or any other data coming through APIs or other means (semi-structured or un-structured).

B. Data Transformation :

Transformation of data from one form to another, its either part of ETL (Extract, Transform and Load) or import/export tools and/or scripts Mainly used to load all sources of data into data processing pipeline. Log management tool can be considered as

part of ETL, as they generate useful events from log files and present dashboard with alerting system in place or they can be directly loaded.

C. Data Processing or Data Integration :

This is yet another source of vast data by combining both structured and un-structured data in one place (either real time or incremental loading); mainly for data processing (Data Warehousing or Analytics) and generates usable data (materialized or aggregated) that can be consumed by data consumption components.

- Hadoop and Ecosystem (Hadoop/HDFS, Map-reduce, HBase, Hive, Impala, Pig etc)- uses HDFS as native storage
- Data Warehouse and Analytics solution (MySQL, SQL Server, Vertica, Green Plum, Aster Data, Exadata, SAP HANA, IBM Netezza, IBM Pure Data, Tera Data, etc.) - Uses vendor specific storage, optionally uses HDFS, even though with degraded performance.
- In-memory Analytics (SAS, Kognitio, Druid, etc.). This is an emerging market and trying to take advantage by reading directly from HDFS. We will see lot of in-memeory analytical in coming days.

D. Data consumption

Data consumption components that either consumes or exposes the data in usable form to end users or to other layers internally (ad-hoc) or externally (using APIs) Reporting

8.6.1 Distributed File System

Day by day storing each day record on the hardware might be quite hectic. Similarly assign a individual form of information over and internet might take a long time period just reading the whole content as the data is roughly stored in the hardware. To store the data properly or essay way in assessing data in as least time as possible, a system known as distributed file system was being introduced; basically system the whole problem is being divided in number of sub-problems (according to related problem is divided). The problems are being processed parallel executing the particular assigned this problem. In the end result of all the sub-problems are being summed up and the result is being supplied to the user.

8.6.2 Big Data and its Importance

Distributed File System has different requirements when compared to that of system local file. These are following the requirements which are to be considered when designing the Distributed File System.

- The fault tolerance feature must be well-implemented. How fast the data can be recovered after any failure becomes one of the most important requirements here.
- Files stored in DFS will be very huge. Most of the files' size exceeds GB level. Handling these types of huge files is very crucial in Distributed File System (DFS). Some file system will divide in number of sub-problems.
- Most of the files in DFS are in write-once-read-many pattern. Therefore many DFS' provide optimized function for file writer and reader. Few of them also have efficient function to edit and arbitrary position in an existing file. Some DFS' don't even provide function to change any existing file.
- Metadata plays a key role in Distributed File System. Since most DFS has the support for millions of files, it's not possible to efficiently retrieve the information on any given file simply by traversing every node directly. Due to this reason, most Distributed File System assigns a certain node as the central, which maintains the metadata of all files stored in the system. The retrieval for file information will become much faster via the metadata list.

Four Vs

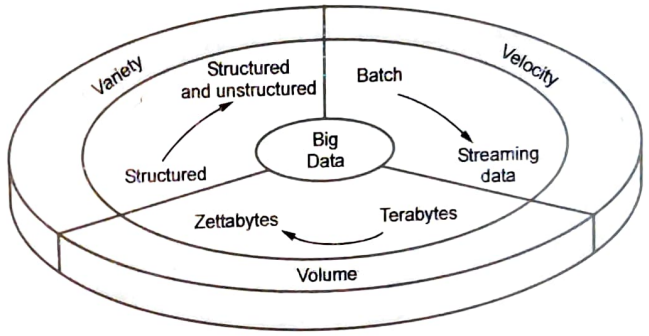


Fig. 8.6.2 Big data

Volume - Volume of the data is its size, and how massive it is. Data volume is primary attribute of big data Data on the earth is growing exponentially due to many reasons. Volume possess the greatest challenge and opportunity as Big Data can help many organization in understanding people better and allocated resources effectively. As well as even then number of records, transactions, tables, or files. Traditional computing methods face difficulties in handling the un-scalable data and its magnitude. The data amount has changed from terabyte to petabyte and now to zettabytes.

Variety - Variety of data to be processed is getting diverse, It is refers to the structural heterogeneity in a dataset. Technological advances allow firms to use various types of structured, semi-structured and unstructured data. Structured data is an organization data in a predefined format. The formatted data resides in fixed fields within a record or file. Unstructured data is a set of data with a complex structure that might or might not have a repeating pattern like e-mails, text, audio, video, or images files etc. While the semi-structured data the combination of both types of data (structured and unstructured). Such type of data does not follow proper structure of data models as in relation database like web data and bibliographic data.

Velocity - Data also raises many issues with the rate with which it is flowing in many organizations exceeding the capacity of the organization systems. The Twitter fire hose (the stream of all tweets, globally, traffic data in mobile communication networks, and streaming Video data are prime examples as this data flows at tremendous rates.

Variability - SAS introduced Variability and complexity as two additional dimensions of big data. Variability refers to the variation in the data flow rates. This factor could be a problem for those who analyze data. As sometime, due to some inconsistency, there may be hampering of data thus creating difficulty in managing and analyzing data.

Value - It is refers to the important feature of the data which is defined by the added-value that the collected data can bring to the intended process, activity or predictive analysis. Big data are often characterized by relatively "low value density". That is, the data received in the original form usually has a low value relative to its volume a high value can be obtained by analyzing large volumes of such data.

Drivers for Big data

Big Data value drivers, that is in what specific ways can Big Data drive economic value with respect to your key business initiatives.

- All transactional data

Access to ALL of your transnational data, Not just 13 months of aggregated data stored in your overly expensive data warehouse, but every customer transaction over the past 10 to 15 years including sales, returns, payments, claims, telephone calls.

- Unstructured data (small data)

Access to internal (consumer comments, work orders, technician notes, service records, email) and external (social media, mobile, census, blogs, newsfeeds) unstructured data that, when combined with your organization's most strategic nouns.

- Low Latency Data Access and Analysis

Low latency data access and analysis to those events where more immediate awareness of that data can lead to faster decisions and actions.

- Predictive Analytics

Integrated predictive analytics to uncover insight, buried in the terabytes of data, across 50 to 70 dimensions and hundreds of metrics

8.6.3 Big Data Analytics

Big Data not only changes the tools one can use for predictive analytics, it also changes our entire way of thinking about knowledge extraction and interpretation. Traditionally, data science has always been dominated by trial-and-error analysis, an approach that becomes impossible when datasets are large and heterogeneous. Ironically, availability of more data usually leads to fewer options in constructing predictive models, because very few tools allow for processing large datasets in a reasonable amount of time. In addition, traditional statistical solutions typically focus on static analytics that is limited to the analysis of samples that are frozen in time, which often results in surpassed and unreliable conclusions. Let's begin with a real world example, looking at a farm that is growing strawberries.

What would a farmer need to consider if they are growing strawberries? The farmer will be selecting the types of plants, fertilizers, pesticides. Also looking at machinery, transportation, storage and labor. Weather, water supply and pestilence are also likely concerns. Ultimately the farmer is also investing the market price so supply and demand and timing of the harvest (which will determine the dates to prepare the soil, to plant, to thin out the crop, to nurture and to harvest) are also concerns.

Let's think about the data available to the farmer, here's a simplified breakdown :

1) Historic weather patterns
2) Plant breeding data and productivity for each Strain
3) Fertilizer specifications
4) Pesticide specifications
5) Soil productivity data
6) Pest cycle data
7) Machinery cost, reliability, fault
8) Water supply data
9) Historic supply and demand data
10) Market spot price and futures data

8.6.4 Big Data Applications

Application of healthcare and medical big data

Big data in the health sector can help doctors make the right choices more quickly, on the basis of information collected by other medical staff. Patients can benefit from more timely and appropriate treatments and be better informed about health care providers. An increased use of data analysis in the health sector can also lead to enormous cost savings through a more precise identification of unnecessary procedures or duplication of tests. Then analysis of large clinical datasets can result in the optimization of the clinical and cost effectiveness of new drugs and treatments.

Market and business

Big Data is the biggest game-changing opportunity for sales and marketing, sicne 20 years ago the Internet went main stream, because of the unprecedented array of insights into customer needs and behaviors. Big data reveals customer's behavior and proven ways to elevate customer experiences. These insights ensure your business's success.

Sports

Sport in business, an increasing volume of information is being collected and captured. Technological advances will fuel exponential growth in this area for the foreseeable future, as athletes are continuously monitored. Statistics can be analyzed and collected to better understand what are the critical factors for optimum performance and success, in all facets of elite sport. Injury prevention, competition, Preparation, and rehabilitation can all benefit by applying this approach. Used consistently this is a powerful measure of progress and performance

Education systems

By using big data analytics in field of education systems, remarkable results can be seen. Data on students online behavior can provide educators with important insights, such as if the course has to be modified or not based on students reception. This modification can be done by making students answer set of online questionnaire and track the accuracy and time taken to answer those questions.

Gaming industry

The amount of data that video game players are generating on a daily basis is growing rapidly. People playing video game and generated lot of data in separate areas: game data, player data and session data. In order to improve their game development, game experience, studios are turning to commercial Hadoop distributions such as Map Reduce to analyze, collect and process dat from these massive data streams. Armed with this valuable insight from big data, video game publishers are now able to enhance

game player engagement and increases player retention by analyzing gamers" social behavior, activity and tracking players" statistics, calculating rewards, quickly generating leader boards, changing game play and mechanics and delivering virtual prizes, so as to creating meaningful gaming experiences for their customers.

Telecommunication industry

Today big challenges for telecommunication are volume, variety and complexity. Telcos combine ETL and traditional relational databases with big data technologies on a single platform. Telcos technology parses, transforms and integrates the vast amount of data generated by location sensors, IPv6 devices, 4G networks and machine to machine monitors: information.Telcos parse and transforms from multiple formats and sources including unstructured mobile, media, web and machine monitor provide data. Telcos masking, managing and identifying sensitive data for regulatory compliance.

Network security

Big data is changing the landscape of security technologies. The tremendous role of big data can be seen in network monitoring. Big data analytics is an effective solution for processing of large scale information as security is major concern in enterprises. Fraud detection is done by using big data analytics. Phone and credit card companies have conducted large-sacle fraud detection for decades. Mainly big data tools are particularly suited to become fundamental for forensics.

Application of banking and financial big data

Data revolution happening in and around 21st century has bound a significance with financial service firms, considering the valuable data they"ve been storing since many decades. And even though the collection of this data was unplanned, since accounting system has always been historical in nature, the potential unlocked by big data analytics exceeds expectation previously expected from this historical record set. This data has now unlocked secrets of money Movements, helped prevent major disasters and thefts and understand consumer behavior. Banks gain the most benefits from big data as they now can extract good information quickly and easily from their data and convert it into meaningful benefits for themselves and their customers.

8.6.5 Tools and Techniques

The tools which are typically used in Big data are as follows

NoSQL

Database MonogoDB, CouchDB, cassandra, Redis, Big Table, Hbase, Hypertable, Voldemort, Riak, ZooKeeper

Map Reduce

Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum

Storage

S3, Hadoop Distributed File System

Servers

EC2, Google App Engine, Elastic, Beanstalk, Heroku

Processing

R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, Big Sheets, Tinkerpop To handle large amount of data various techniques and technologies have been introduced to manipulate, process, analyze and visualize the bigdata.

NoSQL databases

For storing the large of huge data traditional systems were not effective. So, new breed comes into existance that is called NoSQL databases, which are mainaly used to work with big data.

These are non-relational databases.

8.6.6 Map Reduce Approach

A programming task which is divided into multiple identical subtasks and which is distributed among multiple machines for processing is called Map task. The results out of this Map Tasks are combined together into one or many reduce tasks. Overall approach of computing tasks is called a MapReduce approach.

Hadoop

It is openly available software basedon java programming which processes big datasets in a distributed environment. It is designed in such a way if there any hardware failure occures then it automatically handled by the software by the framework. It provides the following parts.

Hadoop Common

Hadoop Distributed File System (HDFS)

Hadoop YARN

Hadoop MapReduce

Hive

Hive runs on the top of Apache hadoop and provides data warehouse capabilities. The Apache Hadoop framework is difficult to understand, and it requires a different approach from traditional programming to write Map Reduce-based programs.

PIG

Another abstraction layer on the top of Map Reduce is provided by Apache Pig. It provides a language called Pig Latin. Pig Latin is a programming language that creates Map Reduce programs using Pig. It is a high-level language for developers to write high level-software for analyzing data. Pig code generates parallel execution tasks, therefore effectively uses the distributed Hadoop cluster.

8.7 Challanges and Opportunities with Big Data

Big data analytics faces different challanges. These are described s follows :

- **Heterogeneity and Incompleteness** : Machine analysis algorithms expect homogeneous data, and cannot understand nuance. Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain.
- **Timeliness** : There are many situations in which the result of the analysis is required immediately. Given a large data set, it is often necessary to find elements in it that meet a specified criterion The larger the data set to be processed, the longer it will take to analyze. It is difficult to design a structure when data is growing in very high speed.
- **Human Collaboration** : A Big Data analysis system must support input from multiple human experts, and shared exploration of results.
- **Privacy and security** : This is another big challange preserving individual privacy. For example in the healthcare industry, record of individual is very personal. But it can be available from multiple sources.
- **Data Quality** : A large volume of data is processed. Analyzing which data is important and to capture it is a big challange.
- **Analysis** : Big data is coming from various data sources. So analytics is a challenge.
- **Skill** : Big data require people with new skill sets. Managing big data effectively requires the right People.

8.7.1 Algorithms using Map Reduce

Map Reduce is a framework for efficiently processing the analysis of big data on a large number of servers. It was developed for the back end of Google's search engine to enable a large number of commodity servers to efficiently process the analysis of huge numbers of web pages collected from all over the world. Apache developed a Project to implement Map Reduce, which was published as open source software (OSS), this enabled many organizations, such as businesses and universities, to tackle big data analysis. It was originally developed by Google and built on well-known principles in

parallel and distributed processing. Since then Map Reduce was extensively adopted for analyzing large data sets in its open source flavor Hadoop.

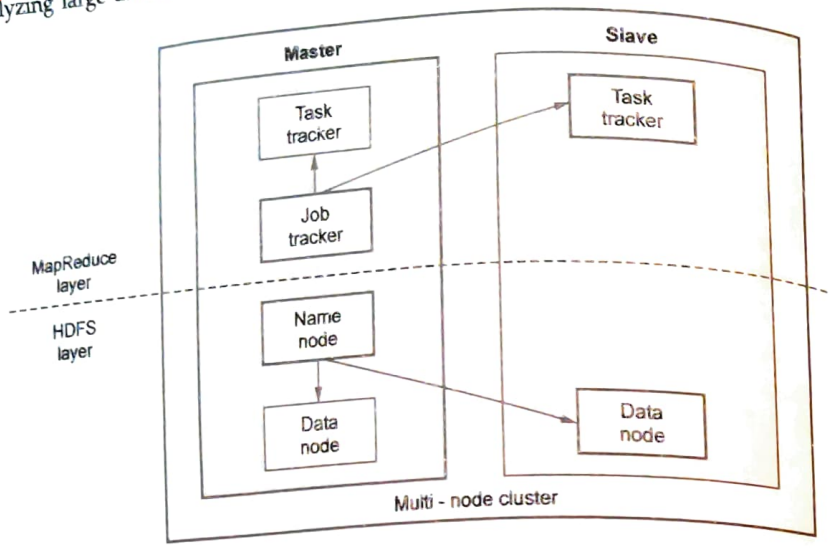


Fig. 8.7.1 AI90 using map reduce

Map Reduce is a simple programming model for processing huge data sets in parallel. Map Reduce have master/slave architecture this is shown in Fig. 8.7.1. The basic notion of Map Reduce is to divide a task into subtasks, handle the sub tasks in parallel, and aggregate the results of the subtasks to form the final output. Programs written in Map Reduce are automatically parallelized : programmers do not need to be concerned about the implementation details of parallel processing. Instead, programmers write two functions : map and reduce. The map phase reads the input (in parallel) and distributes the data to the reducers. Auxiliary phases such as sorting, partitioning and combining values can also take place between the map and reduce phase. Map Reduce programs are generally used to process large files. The input and output for the map and reduce functions are expressed in the form of key value pairs. A Hadoop Map Reduce program also has a component called the Driver. The driver is responsible for initializing the job with its configuration details, specifying the mapper and the reducer classes for the job, informing the Hadoop platform to execute the code on the specified input files(s) and controlling the location where the output files are placed.

In most computation related to high data volumes, it is observed that two main phases are commonly used in most data processing components this is shown in Fig. 8.7.2. Map Reduce created an abstraction phases of Map Reduce model called 'mappers' and 'reducers' (Original idea was inspired from programming languages such

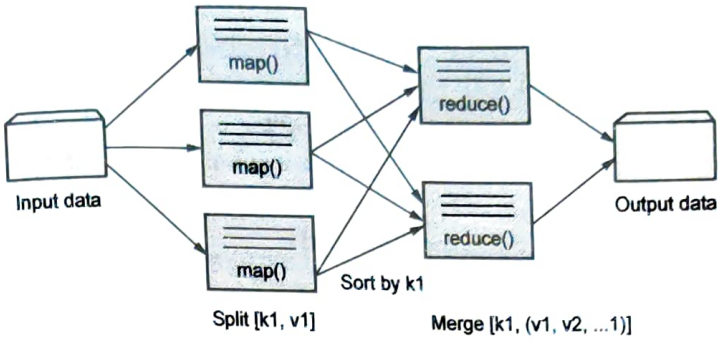


Fig. 8.7.2 Process of map reduce

as Lisp). When it comes to processing large data sets, for each logical record in the input data it is often required to use a mapping function to create intermediate key value pairs. Then another phase called 'reduce' to be applied to the data that shares the same key, to derive the combined data appropriately

Mapper

The mapper is applied to every input key-value pair to generate an arbitrary number of intermediate key-value pairs. The standard representation of this is as follows :

Map (inKey, in Value) -> list (intermediateKey, intermediate Value)

map : (k1, v1) → [(k2, v2)]

The purpose of the map phase is to organize the data in preparation for the processing done in the reduce phase. The input to the map function is in the form of key-value pairs, even though the input to a Map Reduce program is a file or file (s). By default, the value is a data record and the key is generally the offset of the data record from the begning of the data file. The output consists of a collection of key-value pairs which are input for the reduce function. The content of the key-value pairs depends on the specific implementation.

Reducer

The reducer is applied to all values associated with the same intermediate key to generate Output key value pairs.

Reduce (intermediateKey, list(intermediateValue))-> list(outKey, outValue)

Each reduces function processes the intermediate values for a particular key generated by the map function and generates the output. Essentially there exists a one-one mapping between keys and reducers. Several reducers can run in parallel, since they are independent of one another. The number of reducers is decided by the user. By default, the number of reducers is 1 Since we have an intermediate 'group by' operation,

the input to the reducer function is a key value pair where the key-k2 is the one which is emitted from mapper and a list of values [v2] with shares the same key.

Reduce : (K2, [v2]) → [(k3, v3)]

Example 1 : Consider the question of counting the occurrence of each word in the accumulation of large documents. Let"s take 2 input files and perform MapReduce operation on them.

File 1 : bonjour sun hello moon goodbye world

File 2 : bonjour hello goodbye goodluck world earth

Map :

First map : second map

<bonjour,1><bonjour, 1>
<sun,1><hello,1>
<hello,1><goodbye,1>
<moon,1><goodluck,1>
<world,1><earth,1>

Reduce :

<bonjour,2>
<sun,1>
<hello,2>
<moon,1>
<goodluck,1>
<goodbye,2>
<world,2>
<earth,1>

8.7.1 Matrix - Vector Multiplication by Map Reduce

Suppose we have an n×n matrix M, whose element in rows i and column j will be denoted m_{ij} Suppose we also have a vector v of length n, whose j^{th} elements is v_j. Then the matrix-vector product is the Vector x of length n, whose i^{th} element x_i is given by

$$X_i = \sum_{J=1}^{n} m_{ij} v_j$$

If n = 100, we do not want to use a DFS or Map Reduce for this calculation. But this sort of Calculation is at the heart of the ranking of Web pages that goes on at search engines, and there, n is in the tens of billions. Let us first assume that n is large, but not so large that vector v cannot fit in main memory and thus be available to every Map

task. The matrix M and the vector v each will be stored in a file of the DFS. We assume that the row-column Coordinates of each matrix element will be discoverable, either from its position in the file, or because it is stored with explicit coordinates, as a triple (i,j,m_{ij}). We also assume the position of element v_j in the vector v will be discoverable in the analogous way.

The map function

The Map function is written to apply to one element of M. However, if v is not already read into main memory at the compute node executing a Map task, then v is first read, in its entirety, and subsequently will be available to all applications of the Map function performed at this Map task.

Each Map task will operate on a chunk of the matrix M. From each matrix element m_{ij} it produces the key-value pair $(i,m_{ij}v_j)$. Thus, all terms of the sum that make up the component x_i of the matrix-vector product will get the same key, i.

The reduce function :

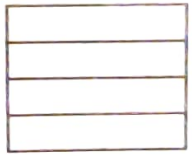
The Reduce function simply sums all the values associated with a given key i. The result will be a pair (i,x_i).

Vector v cannot fit in main memory

However, it is possible for the vector v which is so large to make it fit in its main memory. It is not required that v fit in RAM at a compute node, but if it does not then there will be a very large number of disk accesses as we move pieces of the vector into main memory to multiply components by elements of the matrix. Thus, as an alternative, split the matrix into vertical strips of equivalent width and divide the vector into an equal number of horizontal stripes, of the same height. Our goal is to use enough stripes so that the portion of the vector in one stripe fits conveniently into RAM at a compute node. Fig. 8.7.3 (a) and (b) suggests what the partition looks like if the matrix and vector are each divided into five stripes.

Matrix

Fig. 8.7.3 (a)



Vector

Fig. 8.7.3 (b)

The i^{th} stripe of the matrix multiplies only components from the i^{th} stripe of the vector. Thus, we can divide the matrix into one file for each stripe, and do the same for the vector. Every Map task is allocated a chunk from one of the stripes of the matrix and gets the entire corresponding stripe of the vector. The Map and Reduce tasks can then act exactly as was described above for the case where Map tasks get the entire vector. We shall take up matrix-vector multiplication using Map Reduce. There, because of the particular application (Page Rank calculation), we have an additional constraint that the result vector should be partitioned in the same way as the input vector, so the output may become the input for another iteration of the matrix-vector multiplication.

8.8 Introduction to Hadoop Architecture

Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation. Hadoop makes it possible to run applications on systems with thousands of nodes involving thousands of terabytes. Its distributed file system facilitates rapid data transfer rates among nodes and allows the system to continue operating Uninterrupted in case of a node failure. This approach lowers the risk of catastrophic system failure, even if a significant number of nodes become inoperative.

8.8.1 Hadoop Architecture

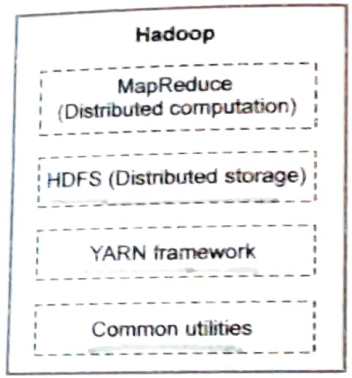


Fig. 8.8.1 Hadoop architecture

- **Hadoop Common :** These are Java libraries and utilities required by other hadoop modules. These libraries provide file system and OS level abstractions and contains the necessary Java files and scripts required to start hadoop.
- **Hadoop YARN :** This is a framework for job scheduling and cluster resource management.

- **Hadoop Distributed File System (HDFS) :** A distributed file system that provides high-throughput access to application data.
- **Hadoop MapReduce :** This is YARN-based system for parallel processing of large data sets.
- **MapReduce :** Hadoop Map Reduce is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner. The term Map Reduce actually refers to the following two different tasks that hadoop programs perform:
- **Map Task :** This is the first task, which takes input data and converts it into a set of data, where individual elements are broken down into tuples.
- **Reduce Task :** This task takes the output from a map task as input and combines those data tuples into a smaller set of tuples. The reduce task is always performed after the map task.

The Map Reduce framework consists of a single master Job Tracker and one slave Task Tracker per cluster-node. The master is responsible for resource management, tracking resource consumption/availability and scheduling the jobs component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves Task Tracker execute the tasks as directed by the master and provide task-status information to the master periodically. The job Tracker is a single point of failure for the Hadoop Map Reduce service which means if, Job Tracker goes down, all running jobs are halted.

8.8.1.1 Hadoop Distributed File System : for (Data Storage)

It has two main parts - a data processing framework and a distributed file system for **data storage.** There's more to it than that, of course, but those two components really make things go. The distributed file system is that far-flung array of storage clusters noted above - i.e., the Hadoop component that holds the actual data. By default, Hadoop uses the cleverly named Hadoop Distributed File System (HDFS), although it can use other file systems as well.

8.8.1.2 Hadoop Distributed File System (HDFS)

The file store is called the Hadoop Distributed File System, or HDFS. HDFS provides scalable, fault-tolerant storage at low cast. The HDFS software detects and compensates for hardware issues, including disk problems and server failure. The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on large clusters of small computer machines in a reliable, fault-tolerant manner. HDFS uses this method when replicating data for data redundancy across multiple racks.

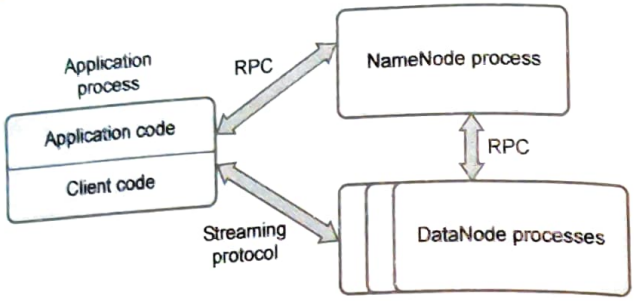


Fig. 8.8.2 Hadoop HDFS

HDFS stores files across a collection of servers in a cluster. Files are decomposed into blocks, and each block is written to more than one (the number is configurable, but three is common) of the servers. This replication provides both fault-tolerance (loss of a single disk or server does not destroy a file) and performance (any given block can be read from one of several servers, improving system throughput.) HDFS ensures data availability by continually monitoring the servers in cluster and the blocks that they manage. Individual blocks include checksums. When a block is read, the checksum is verified, and if the block has been damaged it will be restored from one of its replicas. If a server or disk fails, all of the data is stored is replicated to some other node or nodes in the cluster, from the collection fo replicas. As a result, HDFS runs very well on commodity hardware. It tolerates, and compensate for, failures in the cluster. As clusters get large, even very expensive fault-tolerant servers are likely to fail. Because HDFS expects failure, organizations can spend less on servers and let software compensate for hardware issues.

8.9 Common Hadoop Shell Commands

1. Create a directory in HDFS at given path(s)

Usage :

hadoop fs - mkdir <paths>

Example :

hadoop fs - mkdir / user/ saurzcode/dir1/user/saurzcode/dir2

2. List the contents of a directory

Usage :

hadoop fs - ls <args>

Example :

hadoop fs - ls/user/saurzcode

3. Upload and download a file in HDFS

***Upload* : hadoop fs - put:**

Coply single src file, or multiple src files from local file system to the Hadoop data file system

Usage :

hadoop fs - put <localsrc> ... <HDFS_dest_Path>

Example :

hadoop fs - put/home/saurzcode/Samplefile.txt/user/saurzcode/dir3/

***Download* :**

hadoop fs - get :

Copies/Downloads files to the local file system

Usage :

hadoop fs - get <hdfs_src> <localdst>

Example :

hadoop fs - get/user/saurzcode/dir3/Samplefile.txt/home/

4. See contents of a file

Same as unix cat command :

Usage :

hadoop fs -cat <path[filename]>

Example :

hadoop fs -cat/user/saurzcode/dir1/abc.txt

5. Copy a file from source to destination

This command allows multiple sources as well in which case the destination must be a directory.

Usage :

hadoop fs -cp <source> <dest>

Example :

hadoop fs -cp/user/saurzcode/dir1/abc.txt/user/saurzcode/dir2

6. Copy a file from/To Local file system to HDFS

copy From Local

Usage:

hadoop fs -copy From Local <localsrc> URI

Example :

hadoop fs -copy From Local/home/saurzcode/abc.txt/user/saurzcode/abc.txt

Similar to put command, except that the source is restricted to a local file reference.

copy ToLocal

Usage :

hadoop fs -copy ToLocal [-ignorecrc] [-crc] URI <localdst>

Similar to get command, except that the destination is restricted to a cocal file reference.

7. Move file from source to destination.

Note : Moving files across file system is not permitted

Usage :

hadoop fs -mv <src> <dest>

Example :

hadoop fs -mv/user/saurzcode/dir1/abc.txt/user/saurzcode/dir2

8. Remove a file of dirctory in HDFS.

Remove files specified as argument. Deletes directory only when it is empty

Usage :

hadoop fs -rm <arg>

Example :

hadoop fs -rm /user/saurzcode/dir1/abc.txt

Recursive version of delete.

Usage :

hadoop fs -rmr <arg>

Example :

hadoop fs -rmr/user/saurzcode/

9. Display last few lines of a file.

Similar to tail command in Unix.

Usage :

hadoop fs -tail <path [filename]>

Example:

hadoop fs -tail/user/saurzcode/dir1/abc.txt

10. Display the aggregate length of a file.

Usage :

hadoop fs - du <path>

Example :

hadoop fs -du/user/saurzcode/dir1/abc.txt

8.9.1 Anatomy of File Write and Read

- To start the file read operation, client opens the required file by calling open() on File system object which is an instance of Distributed File System. Open method initiate HDFS client for the read request.
- Distributed File System interacts with Name node to get the block locations of file to be read. Block locations are stored in metadata of name node. For each block, Name node returns the sorted address of Data node that holds the copy of that block. Here sorting is done based on the proximity of Data node with respect to Name node, picking up the nearest Data node first.
- Distributed File System returns an FS Data Input Stream, which is an input stream to support file seeks to the client. FS Data Input Stream uses a wrapper DFS Input Stream to manage I/O operations over Name node and Data node. Following steps are performed in read operation.

a) Client calls read () on DFS Input Stream. DFS Input Stream holds the list of address of block locations on Data node for the first few block of the file. It then locates the first block on closest Data node and connects to it.

b) Block reader gets initialized on target Block/Data node along with below information

- Block ID
- Data start offset to read from
- Length of data to read
- Client name

c) Data is streamed from the Data node back to the client in form of packets, this data is copied directly to input buffer provided by client. DFS client is reading and performing checksum operation and updating the client buffer.

d) Read () is called repeatedly on stream till the end of block is reached. When end of block is reached DFS Input Stream will close the connection to Data node and search next closest Data node to read the block from it.

- Blocks are read in order, once DFS Input Stream done through reading of the first few blocks, it calls the Name node to retrieve Data node locations for the next batch of blocks.
- When client has finished reading it will call Close () on FS Data Input Stream to close the connection.

8.9.2 NameNode, Secondary NameNode and DataNode

NameNode

Name node manages the file system namespace. It maintains the files system tree and the metadata for all the files and directories in the tree. This information is stored persistently on the local disk in the form of two files: the namespace image and the edit log. The namenode also knows the data nodes on which all the blocks for a given file are located; however, it does not store block locations persistently, because this information is reconstructed from data nodes when the system starts.

Secondary NameNode

Secondary NameNode which despite its name does not act as a namenode. Its main role is to periodically merge the namespace image with the edit log to prevent the edit log to prevent the edit log from becoming too large.

DataNode

Datanodes are the workhorses of the filesystem. They store and retrieve blocks when they are told to (by clients or the namenode), and they report back to the namenode periodically with lists of blocks that they are storing.

8.9.2.1 Hadoop Map Reduce Paradigm

Map Reduce is a programming paradigm, developed by Google, which is designed to solve a single problem. It is basically used as an implementation procedure to induce large datasets by using map and reduce operations.

- Normal Log file processing involves processing the log file into reduced data set to load the processed data set into database.
- Computation resources for Normal Log file processing is less.
- Cluster size keeps on increasing with application demand.
- Size of Log File Growing exponentially.
- Log Files of More than 2TB of Size getting generated on daily basis. These log files very hard to process in single machine.
- These files are loaded into HDFS (Hadoop Distributed File System).

- HDFS boon for large files storing and large file processing Map Reduce Programming Patterns used to process the distributed files information.

8.10 Map and Reduce Tasks

- **Lateral computing :** Provides parallel data processing across the nodes of clusters using the Java based API. It works on commodity hardware in case of any hardware failure.
- **Programming languages :** Uses Java, Python and R languages for coding in creating and running jobs for mapper and reducer executables.
- **Data locality :** Ability to move the computational node close to where the data is. That means, the Hadoop will schedule MapReduce tasks close to where the data exist, on which that node will work on it. The idea of bringing the compute to the data rather than bringing data to the compute is the key of understanding MapReduce.
- **Fault tolerant with shared nothing :** The Hadoop architecture is designed where the tasks have no dependency on each other.

8.10.1 Map Reduce Job

The Map Reduce framework operates exclusively on<key,value>pairs, that is, the framework views the input to the job as a set of <key, value> pairs and produces a output of the job as set of <key, value> pairs conceivably of distinct types. The key and value classes have to be serialize able by the framework and hence need to implement the writable interface. A simple Map Reduce program can be written to determine how many times different words appear in a Set of files.

Example 1 : if we the files like file, file 2, and file 3

Input :

File1 : Deer, Bear, River

File2 : Car, Car, River

File3 : Deer, Car, Bear

We can a write a program in map reduce by using three operations like map, combine, reduce

To compute the output

The first step is Mapstep :

First Map	Second Map	Third Map
<Deer,1>	<Car,1>	<Deer,1>

<Bear,1> <Car,1> <Car,1>
<River,1> <River,1> <Bear,1>

The secondary step is Combine Step

<Bear,1><Car,1><Deer,1><River,1>
<Bear,1><Car,1><Deer,1><River,1>
<Car,1>

The final step is ReduceStep :

<Bear,2><Car,3><Deer,2><River,2>

8.11 Task Trackers

Task Tracker executes Mapper/Reducer task as a child process in a separate JVM (Java Virtual Machine). The Child task inherits the environment of the parent Task Tracker. A user can specify environmental variables controlling memory, parallel computation settings, segment size. Requirements of applications using MapReduce specify the Job configuration, input/output locations. It supply map and reduce functions via implementations of appropriate Interfaces and/or abstract classes.

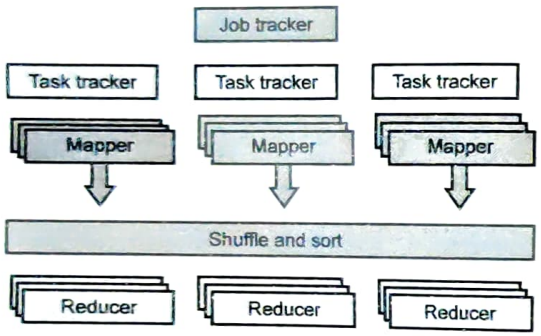


Fig. 8.11.1

8.12 Cluster Setup

Data Clustering deals with partitioning of objects into groups, generally called clusters such that the similarity between members of same group is maximized and similarity between members of different groups is minimized. Various form of distance measure is used to determine similarity of objects.

Purpose

This document describes how to install and configure Hadoop clusters ranging from a few nodes to extremely large clusters with thousands of nodes. To play with Hadoop, you may first want to install it on a single machine.

Prerequisites

- Install Java. See the Hadoop Wiki for known good versions.
- Download a stable version of Hadoop from Apache mirrors.

Installation

Installing a Hadoop cluster typically involves unpacking the software on all the machines in the cluster or installing it via a packaging system as appropriate for your operating system. It is important to divide up the hardware into functions.

Typically one machine in the cluster is designated as the NameNode and another machine as Resource Manager, exclusively. These are the masters. Other services (such as Web App Proxy Server and Map Reduce Job History server) are usually run either on dedicated hardware or on shared infrastructure, depending upon the load.

Operating the Hadoop Cluster

Once the entire necessary configuration is complete, distribute the files to the HADOOP_CONF_DIR directory on all the machines this should be the same directory on all machines. It is recommended that HDFS and YARN run as separate users. In the majority of installations, HDFS processes execute as 'hdfs', YARN is typically using the 'yarn' account.

8.13 SSH and Hadoop Configuration

The hadoop control scripts rely on SSH to perform cluster-wide operations. Hadoop requires SSH access to manage its nodes, remote machines plus your local machine. For our single-node setup of Hadoop, we therefore need to configure SSH access to local host for the Hadoop user we created in the earlier.

The need for SSH Key based authentication is required so that the master node can then login to slave nodes (and the secondary node) to start/stop them, etc. This is also required to be setup on the secondary name node (which is listed in your master's file) so that [presuming it is running on another machine which is a VERY good idea for a production cluster].

8.13.1 HDFS Administering-Monitoring and Maintenance

In the Hadoop world, a **Systems Administrator** is called a Hadoop Administrator. Hadoop Admin Roles and Responsibilities inclu... setting up Hadoop clusters. Other

duties involve backup, recovery and maintenance. Hadoop administration requires good knowledge of hardware systems and excellent understanding of Hadoop architecture.

- Collaborating with application terms to install operating system and Hadoop updates, patches, version upgrades when required.
- Installation and Configuration
- Cluster Maintenance
- Resource Management
- Security Management
- Manage and review Hadoop log files
- HDFS support and maintenance
- Performance tuning of Hadoop clusters and Hadoop MapReduce routines
- Screen Hadoop cluster job performances and capacity planning
- HDFS administration includes monitoring the HDFS file structure, locations, and the update files
- Map Reduce administration includes monitoring the list of applications, configuration of nodes, application status, etc.

Monitoring and Maintenance

HDFS (Hadoop Distributed File System) contains the user directories, input files, and output files. Use the MapReduce commands, put and get, for storing and retrieving.

After starting the Hadoop framework (daemons) by passing the command "start-all.sh" on "/$HADOOP_HOME/..."pass the following URL to the browser "http://localhost:...

Map Reduce Job Monitoring

A MapReduce application is a collection of hobs (Map job, Combiner, Partitioned, and Reduce job). It is mandatory to monitor and maintain the following

- Configuration of data node where the application is suitable.
- The number of data nodes and resources used per application.

8.13.2 The Hadoop Ecosystem

HBases

Hbase is distributed column oriented database where as HDFS is file system. But it is built on top of HDFS sytem. HBase is a management system that is open-source, versioned, and distributed based on the Big Table of Google.

Avro

Avro is data serialization format which brings data interoperability among multiple components of apache hadoop.

Sqoop

Sqoop is tool which can be used to transfer the data from relational database environments like oracle, my sql and post gresql into hadoop environment Sqoop is a command-line interface platform that is used transferring data between relational databases and Hadoop.

Zookeeper

Zookeeper is a distributed coordination and governing service for hadoop cluster It is a centralized service that provides distributed synchronization and providing group services and maintains the configuration information.

Mahout

Mahout is a library for machine-learning and data mining. It is divided into four main groups: collective filtering, categorization, clustering, and mining of parallel frequent patterns.

Review Questions

1. *What is Text mining ?*
2. *Explain Big data with architecture.*
3. *What is map reduce ? Explain.*
4. *What is HDFS ?*
5. *What text mining ? Explain ?*

❑❑❑